


UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

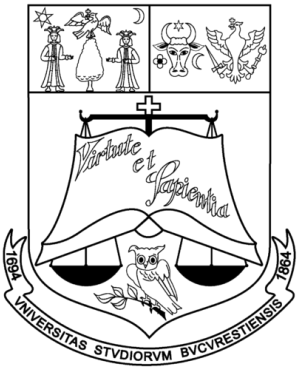
Ph.D. Thesis

# Computational Approaches to Mental Health Disorders Detection from Social Media Texts, Images and Videos

Ana-Maria Bucur

Ph.D. Advisors: Prof. Dr. Liviu P. Dinu, University of Bucharest, Romania
Prof. Dr. Paolo Rosso, Universitat Politècnica de València, Spain

May 2025




UNIVERSITATEA DIN
BUCUREȘTI
VIRTUTE ET SAPIENTIA

Ph.D. Thesis

# Computational Approaches to Mental Health Disorders Detection from Social Media Texts, Images and Videos

Ana-Maria Bucur

Ph.D. Advisors: Prof. Dr. Liviu P. Dinu, University of Bucharest, Romania
Prof. Dr. Paolo Rosso, Universitat Politècnica de València, Spain

May 2025

# Contents

**List of Publications** xv

**List of Figures** xviii

**List of Tables** xx

**1 Introduction** 1
- 1.1 Research Questions . . . 3
- 1.2 Thesis Contributions . . . 6
- 1.3 Thesis Outline . . . 8

**2 Background and Related Work** 9
- 2.1 Mental Health Expression in Language . . . 9
- 2.2 Mental Health Expression in Multimodal Data . . . 10
- 2.3 On the State of NLP Approaches to Mental Disorders Detection in Social Media . . . 10
  - 2.3.1 The Task of Automatic Mental Disorders Detection . . . 11
  - 2.3.2 Shared Tasks for Mental Disorders Identification . . . 13
  - 2.3.3 Early Approaches . . . 14
  - 2.3.4 State-of-the-art . . . 14
  - 2.3.5 Datasets for Depression Detection . . . 17
- 2.4 A Post-COVID-19 Outlook . . . 25
  - 2.4.1 Monitoring Social Media Streams during COVID-19 Pandemic . . . 25
  - 2.4.2 How has the Pandemic Affected the Population? . . . 28
  - 2.4.3 Depression Detection in the Context of the COVID-19 Pandemic . . . 31
- 2.5 Datasets used in this Thesis . . . 33
- 2.6 Summary . . . 36

**3 Language Particularities of Depression in Social Media Data** 37
- 3.1 A Psychologically Informed Part-of-Speech Analysis of Depression in Social Media . . . 38
  - 3.1.1 Part-of-Speech Analysis . . . 38

3.1.2 Results and Discussion . . . 39
3.2 An Exploratory Analysis of the Relation Between Offensive Language and Mental Health . . . 45
3.2.1 Offensive Language Detection and Categorization . . . 45
3.2.2 Results and Discussion . . . 47
3.3 Summary . . . 49

**4 Early Risk Detection of Mental Disorders from Social Media Text 51**
4.1 eRisk 2021: Early Risk Detection of Pathological Gambling and Self-Harm Using BERT . . . 52
4.1.1 Early Detection of Signs of Pathological Gambling . . . 53
4.1.2 Early Detection of Signs of Self-Harm . . . 55
4.2 eRisk 2022: An End-to-End Set Transformer for User-Level Classification of Depression and Gambling Disorders . . . 57
4.2.1 End-to-End Transformer Architecture . . . 58
4.2.2 Interpretability . . . 61
4.2.3 Results . . . 61
4.3 eRisk 2023: Leveraging LLM-generated Data for Detecting Depression Symptoms on Social Media . . . 65
4.3.1 Retrieving Symptoms of Depression . . . 65
4.3.2 Results . . . 69
4.4 Summary . . . 71

**5 Multimodal Depression Detection from Social Media Data 73**
5.1 Depression Detection using Textual and Visual Information . . . 74
5.1.1 Time-Enriched Multimodal Transformers . . . 75
5.1.2 Experiments . . . 77
5.1.3 Results . . . 80
5.2 Depression Detection using Non-Verbal Cues from Videos . . . 85
5.2.1 Multimodal Transformer Architecture for Video Content . . . 85
5.2.2 Experiments . . . 90
5.2.3 Results . . . 93
5.3 Summary . . . 97

**6 The Expression of Happiness in Social Media of Individuals Reporting Depression 98**
6.1 Depression and Expressions of Happiness . . . 101
6.2 Agency and Communion . . . 102
6.3 Extracting Happy Moments from Social Media Data . . . 103
6.4 Inferring Agency and Sociality . . . 110
6.5 Topics of Happy Moments . . . 112
6.6 Sociality and Agency in the Expression of Happy Moments . . . 118
6.7 Discussion . . . 119
6.8 Summary . . . 123

**7 Conclusions** **124**
7.1 Answers to Research Questions . . . . . 124
7.2 Ethical Considerations . . . . . 127
7.3 Conclusions and Future Directions . . . . . 130

# Acknowledgements

I am incredibly grateful to my supervisors, Prof. Liviu Dinu and Prof. Paolo Rosso, for their guidance throughout my Ph.D. journey. Their support made my Ph.D. experience exceptional.

I would like to express my gratitude to Prof. Liviu Dinu for inspiring me to consider pursuing a Ph.D. and for believing in my abilities. When I first met Prof. Dinu, I was intrigued by the application of computational methods for detecting suicidal ideation. At that time, I never imagined that this topic would eventually become part of my Ph.D. research.

I am grateful to Prof. Paolo Rosso for welcoming me to his lab at the Pattern Recognition and Human Language Technologies (PRHLT) Research Center at Universitat Politècnica de València. During my nearly two years in Valencia, I have had the opportunity to experience what it is like to be part of a research lab. I would like to thank my colleagues with whom I shared the lab and research interests: Silvia, Giulia, Betti, Elisa, Eliana, Mariona, Ivan, Damir, Reynier and Angel. I must mention David, who welcomed me from day one and showed me genuine Spanish friendship. I would also like to thank Berta Chulvi, the social psychologist at PRHLT, with whom I have enjoyed a fruitful collaboration.

I would also want to thank the members of the academic guidance and integrity committee, Asst. Prof. Anca Dinu, Assoc. Prof. Traian Rebedea, Prof. Cosima Rughiniș, and Prof. Ioana Podină for their guidance over the years and the valuable feedback they provided throughout my research journey.

I would like to express my deepest appreciation to the two external reviewers, Prof. David Enrique Losada Carril and Assoc. Prof. Arturo Montejo Ráez, for their thorough evaluation of my doctoral thesis.

I would like to express my gratitude to Prof. Ioana Podină for welcoming me to her team of the aiCARE Project. I am also thankful to the colleagues I met there: Diana Todea, Andreea Luca, Liviu Fodor, and Assoc. Prof. Rareș Boian. Throughout the project, I learned how to collaborate and contribute to research with an interdisciplinary team.

I would like to thank Prof. Mihai Dima, the director of the Interdisciplinary School of Doctoral Studies (ISDS), for his encouragement during my Ph.D. I would like to express my gratitude to my colleagues from ISDS: Diana Todea, Andreea Luca, Andreea Codrina-Moldovan, Claudiu Creangă, Marius Micluța-Câmpeanu,

Ștefana Arina Tăbușcă and Andra Florea. Working alongside them during my Ph.D. has provided me with the opportunity for constructive discussions in a truly interdisciplinary environment. Thank you all for your collaboration, support, and friendship.

I am thankful to Prof. Marcos Zampieri, one of my longest collaborators during my Ph.D., for his interest in my topic and for his willingness to teach me how to conduct better research.

I would like to thank Prof. Lucie Flek for welcoming me for a research stay in her Conversational AI and Social Analytics (CAISA) Lab at Philipps-Universität Marburg. I would also like to thank Prof. Hamidreza Jamalabadi for the opportunity to be involved in the Psychiatric Control Systems Lab at the same university.

I am grateful to Petru Soviany for encouraging me during moments when I doubted my ability to conduct research. I also want to thank Ana Sabina Uban for her guidance, especially as I followed her steps by moving to Valencia. I extend my appreciation to my colleagues and friends, Andrei Manolache and Bogdan Iordache, who made my conference experiences enjoyable. I am grateful to my best friend, Cristina, who has been by my side since before we both knew what we wanted to be when we grew up.

I am eternally grateful to my parents, Aurelia and Gheorghe, for always prioritizing my sister's and my education. They went to great lengths to give us, two kids from the countryside, the opportunity to learn and explore the world. I also want to thank my sister, Georgiana, for consistently bringing joy into my life.

Lastly, I am endlessly grateful to Adrian Cosma, my partner in both life and research, for teaching me the joy of conducting "research for fun." Thank you for being by my side during the happy moments of my Ph.D. journey.

# Abstract

Mental disorders affect 1 in 8 people worldwide, with nearly half of the population expected to experience a mental disorder by age 75. Stigma and limited access to care often prevent individuals from seeking the help they need. However, integrating digital technologies can improve accessibility to mental health care. This thesis investigates the use of computational approaches to detect depression, self-harm, and pathological gambling through social media data.

Depression affects approximately 4% of the population and is characterized by various symptoms that can impair a person's ability to function. Self-harm, which mainly impacts young people, involves non-suicidal self-injury practices, while pathological gambling refers to problematic gambling behaviors. Research has shown that specific language patterns in social media data correlate with mental disorders, revealing the potential for using online content in digital screening tools.

This thesis aims to leverage computational models to identify linguistic patterns associated with mental health problems, contributing to early detection and intervention efforts. First, we examine the differences in part-of-speech usage, focusing on self-focused language, pronouns, and verbs, which previous studies have linked to depression. Second, we investigate the prevalence of offensive language, analyzing whether individuals with depression use higher rates of profanity or hate speech. Third, we assess the feasibility of using recent advancements in natural language processing for the early detection of mental disorders, including pathological gambling, self-harm, and depression, by analyzing social media data. In addition, we evaluate the effectiveness of multimodal models that integrate text, images, and video content in detecting depression. Finally, we build upon existing psychological literature, which reveals that despite the challenges of depression, individuals may still experience happiness. Through a large-scale analysis of social media posts, we use a comprehensive deep learning framework to extract, annotate, and classify happy moments based on semantic topics, gender, agency, and sociality.

This thesis highlights the important role of computational approaches in identifying mental health disorders through social media data. By using natural language processing and multimodal models, we can improve early detection and intervention efforts for mental disorders, which have been worsened by the COVID-19 pandemic. Integrating these tools into mental health care can enhance accessibility and support for those who need it.

# Rezumat

Tulburările de sănătate mintală afectează 1 din 8 persoane la nivel global, iar aproape jumătate dintre oameni sunt susceptibili să se confrunte cu o astfel de afecțiune până la vârsta de 75 de ani. Stigmatizarea și accesul limitat la servicii de specialitate reprezintă adesea bariere majore în calea obținerii ajutorului necesar. Cu toate acestea, integrarea tehnologiilor digitale poate contribui semnificativ la îmbunătățirea accesului la îngrijirea în domeniul sănătății mintale. Această teză explorează utilizarea metodelor computaționale pentru detectarea depresiei, a comportamentelor de autovătămare și a dependenței patologice de jocuri de noroc, prin analiza datelor provenite din rețelele sociale.

Depresia afectează aproximativ 4% din populație și se manifestă prin diverse simptome care pot afecta capacitatea de funcționare a unei persoane. Autovătămarea, frecvent întâlnită în rândul tinerilor, implică acte de auto-rănire fără intenții suicidare, în timp ce jocul patologic face referire la comportamente disfuncționale legate de jocurile de noroc. Studiile recente au arătat că anumite tipare lingvistice identificate în postările de pe rețelele sociale sunt corelate cu diverse tulburări mintale, ceea ce evidențiază potențialul folosirii conținutului generat de utilizatori în instrumentele digitale de screening.

Scopul principal al acestei teze este de a utiliza modele computaționale pentru identificarea tiparelor de limbaj asociate cu probleme de sănătate mintală, contribuind astfel la detectarea timpurie și la intervenții mai eficiente. În primul rând, sunt analizate diferențele în utilizarea părților de vorbire, cu accent pe limbajul autoreferențial, pronume și verbe, elemente asociate în literatura de specialitate cu depresia. În al doilea rând, este evaluată frecvența limbajului ofensator, urmărindu-se dacă persoanele afectate de depresie recurg mai des la injurii sau discursuri de ură. În al treilea rând, se analizează potențialul recentelor progrese în procesarea limbajului natural în identificarea precoce a tulburărilor precum jocul patologic, autovătămarea și depresia, prin intermediul datelor extrase din rețelele sociale. Lucrarea evaluează totodată eficiența modelelor multimodale, care combină text, imagini sau conținut video, în detectarea simptomelor depresive. În cele din urmă, bazându-ne pe literatura din psihologie ce evidențiază faptul că, în ciuda suferinței asociate depresiei, persoanele pot continua să trăiască momente de fericire, ne propunem să analizăm la scară largă postărilor din rețelele sociale. Prin folosirea algoritmilor de procesarea limbajului natural, se extrag, adnotează și clasifică

momentele fericite în funcție de tematică semantică, gen, agenție și dimensiunea socială.

Această teză evidențiază rolul important al metodelor computaționale în identificarea tulburărilor mintale prin intermediul datelor din rețelele sociale. Aplicarea procesării limbajului natural și a modelelor multimodale poate îmbunătăți semnificativ eforturile de prevenție și intervenție timpurie, într-un context în care impactul pandemiei de COVID-19 a accentuat nevoia unor soluții accesibile și eficiente în domeniul sănătății mintale.

# Resumen

Los trastornos de la salud mental afectan a 1 de cada 8 personas en todo el mundo, y se estima que casi la mitad de la población experimentará un trastorno mental antes de los 75 años. El estigma y el acceso limitado a la atención médica a menudo impiden que las personas busquen la ayuda que necesitan. Sin embargo, la integración de tecnologías digitales puede mejorar la accesibilidad a la atención en salud mental. Esta tesis investiga el uso de enfoques computacionales para detectar depresión, autolesiones y ludopatía patológica a través de datos provenientes de redes sociales.

La depresión afecta aproximadamente al 4% de la población y se caracteriza por diversos síntomas que pueden perjudicar la capacidad funcional de una persona. Las autolesiones afectan especialmente a los jóvenes e implican prácticas de autolesión no suicida, mientras que la ludopatía patológica se refiere a comportamientos problemáticos relacionados con el juego. Las investigaciones han demostrado que ciertos patrones lingüísticos en publicaciones de redes sociales se correlacionan con trastornos mentales, lo que revela el potencial del contenido generado por los usuarios como herramienta de detección.

Esta tesis tiene como objetivo aprovechar modelos computacionales para identificar patrones lingüísticos asociados con problemas de salud mental, contribuyendo a los esfuerzos de detección temprana e intervención. En primer lugar, se examinan las diferencias en el uso de categorías gramaticales, enfocándose en el lenguaje centrado en uno mismo, los pronombres y los verbos, los cuales han sido vinculados a la depresión en estudios previos. En segundo lugar, se investiga la prevalencia del lenguaje ofensivo, analizando si las personas con depresión utilizan tasas más altas de blasfemias o discursos de odio. En tercer lugar, se evalúa la viabilidad de utilizar los avances recientes en procesamiento del lenguaje natural para la detección temprana de trastornos mentales, incluyendo la ludopatía, las autolesiones y la depresión, a partir del análisis de datos en redes sociales. Además, se evalúa la eficacia de modelos multimodales que integran texto, imágenes y contenido en video para detectar depresión. Finalmente, se amplía el conocimiento basado en la literatura psicológica existente, que revela que, a pesar de los desafíos de la depresión, las personas aún pueden experimentar felicidad. A través de un análisis a gran escala de publicaciones en redes sociales, se utiliza un marco integral de aprendizaje profundo para extraer, anotar y clasificar momentos felices según temas

semánticos, género, agencia y sociabilidad.

Esta tesis resalta el papel importante de los enfoques computacionales para identificar trastornos de salud mental mediante datos de redes sociales. Al utilizar procesamiento del lenguaje natural y modelos multimodales, es posible mejorar la detección temprana y los esfuerzos de intervención frente a los trastornos mentales, que se han agravado con la pandemia de COVID-19. Integrar estas herramientas en la atención en salud mental puede mejorar el acceso y el apoyo para quienes lo necesitan.

# Resum

Els trastorns mentals afecten una de cada huit persones arreu del món, i s'estima que quasi la mitat de la població experimentarà algun trastorn mental abans dels 75 anys. L'estigma i l'accés limitat a l'atenció sovint impedeixen que les persones busquen l'ajuda que necessiten. Tanmateix, la integració de tecnologies digitals pot millorar l'accessibilitat a l'atenció en salut mental. Aquesta tesi investiga l'ús d'enfocaments computacionals per detectar la depressió, l'autolesió i el joc patològic a través de dades extretes de les xarxes socials.

La depressió afecta aproximadament el 4% de la població i es caracteritza per diversos símptomes que poden afectar la capacitat funcional de la persona. L'autolesió impacta especialment els joves i implica pràctiques d'autolesió no suïcida, mentre que el joc patològic es refereix a comportaments problemàtics relacionats amb el joc. La recerca ha demostrat que certs patrons lingüístics en publicacions de xarxes socials es correlacionen amb trastorns mentals, la qual cosa mostra el potencial del contingut generat per usuaris com a eina de detecció.

Aquesta tesi té com a objectiu aprofitar models computacionals per identificar patrons lingüístics associats amb problemes de salut mental, contribuint a la detecció precoç i a les intervencions. En primer lloc, s'analitzen les diferències en l'ús de les categories gramaticals, centrant-se en el llenguatge autoreferencial, els pronoms i els verbs, que han estat relacionats amb la depressió en estudis previs. En segon lloc, s'investiga la prevalença del llenguatge ofensiu, analitzant si les persones amb depressió utilitzen més blasfèmies o discursos d'odi. En tercer lloc, s'estudia la viabilitat d'utilitzar els avanços recents en el processament del llenguatge natural per a la detecció precoç de trastorns mentals, incloent-hi el joc patològic, l'autolesió i la depressió, mitjançant l'anàlisi de dades de xarxes socials. A més, s'avalua l'eficàcia dels models multimodals que integren text, imatges i contingut audiovisual per detectar la depressió. Finalment, es construeix a partir de la literatura psicològica existent, que mostra que, tot i els desafiaments de la depressió, les persones encara poden experimentar felicitat. A través d'un anàlisi a gran escala de publicacions en xarxes socials, s'utilitza un marc complet d'aprenentatge profund per extraure, anotar i classificar moments feliços segons temes semàntics, gènere, agència i sociabilitat.

Aquesta tesi destaca el paper clau dels enfocaments computacionals en la identificació de trastorns mentals mitjançant dades de xarxes socials. Amb l'ús del

processament del llenguatge natural i models multimodals, es poden millorar els esforços de detecció precoç i d'intervenció davant trastorns mentals, que s'han vist agreujats per la pandèmia de la COVID-19. Integrar aquestes eines en l'atenció en salut mental pot millorar l'accessibilitat i el suport per a les persones que ho necessiten.

# List of Publications

1. **A.-M. Bucur** and L. P. Dinu, *Detecting Early Onset of Depression From Social Media Text Using Learned Confidence Scores*, Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it) (2020)

2. **A.-M. Bucur**, A. Cosma and L. P. Dinu, *Early Risk Detection of Pathological Gambling, Self-Harm and Depression Using BERT*, Working Notes of CLEF 2021, Early Risk Prediction on the Internet Lab (eRisk), CEUR 2936 (2021) pp. 938–949

3. **A.-M. Bucur**, I. R. Podina and L. P. Dinu, *A Psychologically Informed Part-of-Speech Analysis of Depression in Social Media*, Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP) (2021) pp. 199–207

4. **A.-M. Bucur**, M. Zampieri and L. P. Dinu, *An Exploratory Analysis of the Relation Between Offensive Language and Mental Health*, Findings of the Association for Computational Linguistics: ACL-IJCNLP (2021) pp. 3600–3606, *CORE A* Conference*

5. **A.-M. Bucur**, A. Cosma and L. P. Dinu, *Sequence-to-Sequence Lexical Normalization With Multilingual Transformers*, Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT), EMNLP (2021) pp. 473–482

6. **A.-M. Bucur**, H. Jang and F. F. Liza, *Capturing Changes in Mood Over Time in Longitudinal Data Using Ensemble Methodologies*, Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology (CLPsych), NAACL (2022) pp. 205–212

7. **A.-M. Bucur**, A. Cosma, L. P. Dinu and P. Rosso, *An End-to-End Set Transformer for User-Level Classification of Depression and Gambling Disorder*, Working Notes of CLEF 2022, Early Risk Prediction on the Internet Lab (eRisk), CEUR 3180 (2022) pp. 851–863

8. **A.-M. Bucur**, A. Cosma and I.-B. Iordache, *BLUE at Memotion 2.0 2022: You Have My Image, My Text and My Transformer*, Proceedings of the

First Workshop on Multimodal Fact-Checking and Hate Speech Detection (DE-FACTIFY), AAAI (2022)

9. **A.-M. Bucur**, A. Cosma and L. P. Dinu, *Life Is Not Always Depressing: Exploring the Happy Moments of People Diagnosed With Depression*, Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC) (2022) pp. 4186–4192, *CORE C Conference*

10. **A.-M. Bucur**, *Utilizing ChatGPT Generated Data to Retrieve Depression Symptoms From Social Media*, Working Notes of CLEF 2023, Early Risk Prediction on the Internet Lab (eRisk), CEUR 3497 (2023) pp. 662–671

11. I. R. Podina, **A.-M. Bucur**, D. Todea, L. Fodor, A. Luca, L. P. Dinu and R. F. Boian, *Mental Health at Different Stages of Cancer Survival: A Natural Language Processing Study of Reddit Posts*, Frontiers in Psychology 14 (2023) pp. 1150227, *IF 2.6, Q2 Journal*

12. I. R. Podina, **A.-M. Bucur**, L. Fodor and R. Boian, *Screening for Common Mental Health Disorders: A Psychometric Evaluation of a Chatbot System*, Behaviour & Information Technology (2023), *IF 2.9, Q2 Journal*

13. **A.-M. Bucur**, A. Cosma, P. Rosso and L. P. Dinu, *It's Just a Matter of Time: Detecting Depression With Time-Enriched Multimodal Transformers*, Advances in Information Retrieval. ECIR 2023. Lecture Notes in Computer Science 13980 (2023) pp. 200–215, *CORE A Conference*

14. D. Gimeno-Gómez, **A.-M. Bucur**, A. Cosma, C.-D. Martínez-Hinarejos and P. Rosso, *Reading Between the Frames: Multi-Modal Depression Detection in Videos From Non-Verbal Cues*, Advances in Information Retrieval. ECIR 2024. Lecture Notes in Computer Science 14608 (2024) pp. 191–209, *CORE A Conference*

15. V. Varadarajan, A. Lahnala, G. A. V, G. Dey, S. Mangalik, **A.-M. Bucur**, N. Soni, R. Rao, K. Lanning, I. Vallejo, L. Flek, H. A. Schwartz, C. Welch and R. L. Boyd, *Archetypes and Entropy: Theory-Driven Extraction of Evidence for Suicide Risk*, Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych), EACL (2024) pp. 278–291

16. M. N. Raihan, S. S. Chowdhury Puspo, M. S. Rahman Farabi, **A.-M. Bucur**, T. Ranasinghe and M. Zampieri, *MentalHelp: A Multi-Task Dataset for Mental Health in Social Media*, Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING) (2024) pp. 11196–11203, *CORE B Conference*

17. R. Boian, **A.-M. Bucur**, D. Todea, A. I. Luca, T. Rebedea and I. Podina, *A Conversational Agent Framework for Mental Health Screening: Design,*

*Implementation, and Usability*, Behaviour & Information Technology (2024), *IF 2.9, Q2 Journal*

18. **A.-M. Bucur**, *Leveraging LLM-Generated Data for Detecting Depression Symptoms on Social Media*, Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2024. Lecture Notes in Computer Science 14958 (2024) pp. 193–204

19. **A.-M. Bucur**, B. Chulvi, A. Cosma and P. Rosso, *The Expression of Happiness in Social Media of Individuals Reporting Depression*, IEEE Transactions on Affective Computing 16.1 (2025) pp. 360–375, *IF 9.6, Q1 Journal*

20. **A.-M. Bucur**, A.-C. Moldovan, K. Parvatikar, M. Zampieri, A. R. KhudaBukhsh and L. P. Dinu, *On the State of NLP Approaches to Modeling Depression in Social Media: A Post-Covid-19 Outlook*, IEEE Journal of Biomedical and Health Informatics (2025), *IF 6.7, Q1 Journal*

21. **A.-M. Bucur**, A.-C. Moldovan, K. Parvatikar, M. Zampieri, A. R. KhudaBukhsh and L. P. Dinu, *Datasets for Depression Modeling in Social Media: An Overview*, Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych), NAACL (2025)

22. A. Sandu, T. Mihailescu, A. S. Uban and **A.-M. Bucur**, *Capturing the Dynamics of Mental Well-Being: Adaptive and Maladaptive States in Social Media*, Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych), NAACL (2025)

23. **A.-M. Bucur**, M. Zampieri, T. Ranasinghe and F. Crestani, *A Survey on Multilingual Mental Disorders Detection From Social Media Data*, Pre-Print (2025) (Submitted)

24. N. Raihan, **A.-M. Bucur** and M. Zampieri, *Large Language Models for Mental Health: A Multilingual Evaluation*, Pre-Print (2025) (Submitted)

25. **A.-M. Bucur**, T. A. Koosha, A. Cosma, L. Flek, S. E. Thanarajah, F. Bernhard, P. Rosso and H. Jamalabadi, *Vanishing Boundaries: A Unifying Account of Multidimensional Emotion Dynamics and Alterations in Depression*, Pre-Print (2025) (Submitted)

# List of Figures


2.1 Number of papers on depression modeling published each year in peer-reviewed conferences or journals . . . 11
2.2 Overview of tasks related to detecting mental health problems from social media. . . . 12
2.3 The most used datasets for depression detection . . . 18
2.4 The most used platforms for collecting mental-health related data . 19
2.5 Taxonomy of papers published during the pandemic (2020-2023) . 26

3.1 Frequency of content part-of-speech. . . . 40
3.2 Most frequent nouns and verbs. . . . 41
3.3 Frequency of verb tenses. . . . 41
3.4 Frequency of functional part-of-speech. . . . 42
3.5 Frequency of personal pronouns. . . . 43
3.6 Indices based on part-of-speech tags. . . . 43
3.7 Feature importances of the RandomForest model for depression classification using POS frequencies and POS indices . . . 44
3.8 Distribution of the offensive language score for posts written by users with self-reported depression diagnosis . . . 47
3.9 Keyness for words from posts showing/not showing signs of depression 49

4.1 Proposed model architecture for eRisk 2022 . . . 59
4.2 Performance of our model across training steps, in terms of $F_1$ score, for different sentence encoders (RoBERTa / MiniLM) . . . 60
4.3 Performance of our model across training steps, in terms of validation $F_1$ score, for RoBERTa sentence embeddings and varying the $K$, the number of texts per user . . . 60
4.4 Texts from a particular user, ranked in terms of attribution scores . 61
4.5 Comparison of most frequent words between the original BDI-II queries, generated queries and the manually annotated relevant documents. . . . 71

5.1 The overall architecture of our proposed method for multimodal depression detection from images and texts . . . 75
5.2 The two methods used for sampling posts from a user's timeline . 76

5.3 Distribution of posts per user and the average time between the posts on Reddit and Twitter multimodal datasets . . . . . 78
5.4 Comparison among different window sizes used in our experiments, using EmoBERTa for text embeddings and CLIP for image encodings 82
5.5 Error analysis on two predictions, one correct, and another incorrect 83
5.6 The overall architecture of our proposed method for depression detection using non-verbal cues . . . . . 86
5.7 Illustrative example of fractional positional embedding for temporally aligning video and audio sampling rates . . . . . 88
5.8 Distributions of video durations for each of our benchmarking datasets: D-Vlod, DAIC-WOZ and E-DAIC . . . . . 91
5.9 Presence distributions for each of our considered modalities in D-Vlog 91
5.10 Attribution scores per each modality across frames obtained with Integrated Gradients on a selected window from a subject suffering from depression from D-Vlog . . . . . 96

6.1 Diagram of our method for extracting happy moments from controls and individuals with mentions of depression diagnoses from their social media posts . . . . . 104
6.2 Topic clusters of happy moments represented using supervised dimensionality reduction with UMAP . . . . . 114
6.3 Agency and Sociality differences in users from the depression and control group by gender with confidence intervals . . . . . 119

# List of Tables


2.1 List of available datasets for depression detection using data posted on online platforms . . . 21

3.1 Percentage of posts labeled as offensive from total posts of self-reported individuals . . . 47
3.2 Percentage of posts labeled as a targeted insult from the offensive posts of individuals with self-reported depression diagnoses . . . 48

4.1 Decision-based evaluation on Task 1 of eRisk 2021: Early Detection of Signs of Pathological Gambling . . . 54
4.2 Ranking-based evaluation on Task 1 of eRisk 2021: Early Detection of Signs of Pathological Gambling . . . 54
4.3 Decision-based results for Task 2 of eRisk 2021: Early Detection of Signs of Self-Harm . . . 55
4.4 Ranking-based evaluation for Task 2 of eRisk 2021: Early Detection of Signs of Self-Harm . . . 56
4.5 Decision-based evaluation on Task 1 of eRisk 2022: Early Detection of Signs of Pathological Gambling . . . 63
4.6 Ranking-based evaluation on Task 1 of eRisk 2022: Early Detection of Signs of Pathological Gambling . . . 63
4.7 Decision-based evaluation on Task 2 of eRisk 2022: Early Detection of Depression . . . 64
4.8 Ranking-based evaluation on Task 2 of eRisk 2022: Early Detection of Depression . . . 64
4.9 Instructions for the data generation that ensure diversity and adherence to the BDI-II responses. . . . 67
4.10 Examples generated by GPT 3.5 for the first symptom of BDI-II: Sadness . . . 68
4.11 Examples generated by LLama-3 for the second symptom of BDI-II: Pessimism. . . . 68
4.12 Ranking-based evaluation (majority voting) for Task 1: Search for symptoms of depression . . . 69

4.13 Ranking-based evaluation (unanimity) for Task 1: Search for symptoms of depression . . . 69

5.1 Statistics for the Reddit [64] and Twitter [163] datasets . . . 78
5.2 Results for multimodal depression detection on Twitter dataset [163] 81
5.3 Results for multimodal depression detection on multiRedditDep . 81
5.4 Model comparison with different text and image encoding methods using the Reddit and Twitter datasets, with time2vec positional embeddings and 128 window size . . . 83
5.5 Results on the D-Vlog test set compared to existing approaches . . 93
5.6 Performance of our multimodal method on the DAIC-WOZ test set 94
5.7 Performance of our multimodal method on D-Vlog using different modality combinations, compared to the Perceiver . . . 94
5.8 Performance of our multimodal method on the E-DAIC test set . . 95

6.1 Overview of the datasets used for happy moments analysis. . . . 105
6.2 Results for happy moments extraction using Positive-Unlabeled Learning . . . 108
6.3 Examples of extracted happy moments and their corresponding context . . . 109
6.4 Results for semi-supervised classification of agency and sociality from CL-Aff dataset . . . 112
6.5 Selected examples of happy moments with their corresponding predictions for agency and sociality . . . 112
6.6 Gender per topic for happy moments . . . 116
6.7 Happy moments of depressed and control users by gender . . . 117
6.8 Agency and Sociality differences in users from the depression and control group using Chi-square test . . . 118

CHAPTER 1

# Introduction

Mental disorders are a public health concern, affecting one in eight individuals around the world.[1] It is estimated that nearly half of the population will experience at least one mental disorder by the age of 75 [26]. Unfortunately, many individuals do not seek psychiatric help for mental health issues due to stigma. Stigma varies across cultures and is influenced by cultural norms, religious beliefs, and social attitudes [27]. Due to the stigma associated with mental health and limited access to professional care—particularly in low-income countries—the World Health Organization advocates for improved delivery of mental health services.[2] This includes the integration of digital technologies to enhance accessibility and support. Digital technologies can improve mental health care by providing screening tools through social media or smartphones [12, 17] and delivering culturally adapted digital interventions [28]. In this thesis, we focus on the use of computational approaches for the detection of depression, self-harm and pathological gambling using social media data. While these computational approaches can be valuable for the initial screening of mental disorders, they should only be used within a human-in-the-loop framework. We emphasized the interpretability of our methods and developed models that provide insight into the predictions made to assist experts in their practice.

Depression is one of the most prevalent mental disorders, affecting approximately 4% of the global population.[3] "Depression is a disorder of mood that is characterized by a triad of negative views: about the self, the world, and the future." [29]. Depression is characterized by symptoms such as depressed mood, loss of interest in activities, changes in body weight, sleep disturbances, fatigue, psychomotor agitation or retardation, feelings of guilt, and thoughts of death or suicidal ideation [30]. For individuals to be diagnosed with depression, symptoms must be persistent

[1] https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[2] https://www.who.int/news/item/17-06-2022-who-highlights-urgent-need-to-transform-mental-health-and-mental-health-care

[3] https://www.who.int/news-room/fact-sheets/detail/depression

and significantly impair their ability to function. Each person's experience with depression is unique, with distinct symptoms and experiences. For instance, the way people perceive time can be significantly affected by depression. Some individuals may feel as though time is slowing down, while others may have a sense that time is running out. In addition, other individuals report that their perception of time remains unaffected by depression [31]. William Styron, in his memoir about his experience with depression, emphasizes that depression is complex and each person's experience is unique. He notes that, while there are some common characteristics of depression, he was "amazed at some of the freakish phenomena—not reported by other patients—that it has wrought amid the twistings of my mind's labyrinth." [32]

Self-harm or nonsuicidal self-injury disorder is a mental health condition characterized by self-inflicting destruction of body tissue without suicidal intent [33]. Adolescents and young adults are particularly at risk of engaging in self-harming behavior, with the overall lifetime prevalence of nonsuicidal self-injury disorder estimated to be almost 17% [34].

Pathological gambling is a mental health disorder characterized by persistent and problematic gambling behavior [30]. Individuals with gambling addiction are constantly preoccupied with gambling, spending increasing amounts of money, and may try to hide their problems from others. Although they may try to control or eliminate their addiction, they often face multiple unsuccessful attempts. The growth of Internet access has significantly increased the availability of gambling opportunities, which has contributed to a rise in the incidence and prevalence of pathological gambling [35, 36].

Research from psychology indicates that there are language patterns associated with mental disorders. For instance, a tendency towards self-focused attention is reflected in language through first-person singular pronouns [37]. The writings of individuals with suicide ideation tend to contain more negative and dichotomous words (e.g., "never," "always") [38]. These language patterns can be identified from clinical interviews [39] and short essay writings [37]. Similarly, these patterns can also be detected from social media data. De Choudhury et al. [40] conducted the first work on detecting mental disorders using social media, and analyzed Twitter data to identify individuals with depression. Their depression status was determined based on results from the CES-D screening questionnaire.[4] Other studies have indicated that language used on Facebook can predict future depression diagnoses recorded in medical records, suggesting that social media data could be a valuable complement to traditional depression screening methods [41]. Moreover, Coppersmith et al. [42] analyzed social media language before a suicide attempt and showed that there are quantifiable signals that can be used for suicide risk assessment. The vast amount of user-generated content available on social media enables researchers to apply computational models to explore signs and symptoms of various mental health conditions using online data.

[4] Center for Epidemiologic Studies Depression Scale: https://www.apa.org/depression-guideline/epidemiologic-studies-scale.pdf

Research has shown promising results in predicting various mental disorders through social media analysis and natural language processing (NLP) models [43, 44]. Social media platforms are a valuable data source, facilitating low-cost and non-invasive early interventions. According to Amir et al. [45], mental health estimates derived from social media data correlate with statistics obtained through traditional methodologies from the US National Institute of Mental Health. This suggests that using social media could be a promising approach to complement existing epidemiological practices. Skaik and Inkpen [46] have used social media data annotated by self-mentions of depression diagnoses to build a model for depression prediction. This model was used to predict the prevalence of depression in a Canadian population sample. The results were consistent with the official depression statistics released by the Canadian Community Health Survey (CCHS).

In this thesis, we focused on analyzing data from individuals who have received a formal diagnosis of mental disorders. Since obtaining this information can be challenging, we primarily relied on available datasets in which users' specific mentions of mental health diagnoses from social media data were considered as a proxy for formal diagnoses. We explore the particularities of online language from individuals with depression in terms of part-of-speech tags and the use of offensive language. We develop models for the detection of depression, self-harm and pathological gambling, obtaining competitive results on benchmark datasets from the eRisk Lab.[5] In addition, we develop multimodal models that include text and image data or non-verbal cues from videos to detect depression. Lastly, we investigate the expression of happiness in social media data from individuals with self-reported depression diagnoses.

## 1.1 Research Questions

In this thesis, we aim to address the following Research Questions (RQs):

**RQ1 Are there significant differences in part-of-speech usage in social media texts between individuals with a self-reported depression diagnosis and a control group?**

Previous research has highlighted some particularities of language use in individuals diagnosed with depression. For instance, individuals with depression tend to use more self-focused language, as evidenced by the increased use of the personal pronoun *"I"* [37, 38]. Additionally, they show cues of rumination, by using more past tense verbs [47]. Building on these findings, we aim to conduct a comprehensive analysis of part-of-speech use in social media text of individuals with self-reported depression diagnosis compared to a control group. We seek to determine if the patterns found in previous research are also present in social media texts. We address this question in

[5] https://erisk.irlab.org/

Section 3.1, by analyzing social media data from Reddit and monitoring the use of universal part-of-speech tags, verb tenses, and the person and number of pronouns [3].

**RQ2 Are there significant differences in the use of offensive language between individuals with self-reported depression diagnoses and control?**

Towards uncovering novel insights into the expression of depression in social media, we investigate the use of offensive language content by individuals with self-mentions of depression diagnoses. This exploration is motivated by findings from Birnbaum et al. [48], which have shown that users with mood disorders and schizophrenia spectrum disorders use more swear words in their Facebook messages compared to control users. In Section 3.2, we build on recent research in offensive language identification and apply it to mental health datasets. We analyze the use of offensive language in social media texts of individuals with self-reported depression diagnoses and control, using two publicly available datasets sourced from Reddit. Additionally, we investigate whether the offensive language is targeted toward others, in the form of hate speech, or if it consists merely of profanity [4].

**RQ3 How have recent advancements in artificial intelligence and natural language processing improved the early detection of mental disorders such as pathological gambling, self-harm, and depression, from social media texts?**

For many people living with mental disorders, accessing treatment or seeking help can be challenging because of social stigma, and many individuals remain undiagnosed. Early risk detection of mental illnesses has a massive positive impact on the well-being of people, as it can lead to early intervention and treatment. The Early Risk Prediction on the Internet Lab (eRisk) Lab, part of the Conference and Labs of the Evaluation Forum (CLEF), focuses on the effectiveness and practical applications of early risk detection of various mental disorders. Each year, the eRisk Lab organizes tasks for the early risk detection of depression, pathological gambling, self-harm, or retrieving symptoms of depression from social media text. Pathological gambling is a mental health illness in which a person has persistent and repetitive problematic gambling behavior. Deliberate self-harm refers to the act of intentionally inflicting harm on oneself, resulting in tissue damage. Depression is a mood disorder, which is the the leading cause of disability globally. In Chapter 4, we present our proposed methods based on the recent advances in natural language processing for the early risk detection of mental disorders from social media text. We analyze the performance of the models based on our participation in the tasks organized by the eRisk Lab [2, 7, 10, 18].

**RQ4 Are multimodal models more effective than unimodal models in detecting**

**depression from social media data that includes text, images, or video content?**

Although unimodal models have demonstrated good performance in detecting depression from social media text [2, 49, 50], multimodal models have been shown to outperform them [51–53]. We propose to evaluate the effectiveness of multimodal models in two different contexts. In Section 5.1, we present the development of a multimodal model that incorporates texts and images from two different social media platforms, Reddit and Twitter [13]. The inclusion of images in our model is motivated by the fact that images can provide additional insights into a user's mental state. For instance, Reece and Danforth [54] found that images posted online by individuals with depression tend to be sadder and have bluer, darker, and grayer tones compared to those from healthy individuals. Furthermore, users with depression post more images featuring faces but include fewer faces per image, indicating reduced social interactivity and increased self-focus [54, 55].

In Section 5.2, we present the development of a multimodal model that relies on visual and auditory information from videos, without incorporating textual information. Rather than analyzing language patterns, our approach focuses on patterns derived from facial expressions, body language, and vocal characteristics. Depression is often characterized by psychomotor changes, such as agitation or retardation, which are closely linked to other symptoms, such as depressed mood, sleep disturbances, changes in appetite, and a loss of interest in activities. Some psychomotor particularities associated with depression include reduced facial expressiveness [56], slower speech rate, longer pause duration [57], and downward head tilt [58]. We propose a multimodal temporal model that processes various non-verbal cues across time in order to estimate depression from both noisy, in-the-wild videos and from controlled laboratory recordings [14].

**RQ5 How can deep learning models identify happy moments from the social media data of individuals with self-reported depression diagnoses? Are there differences in these happy moments compared to a control group?**

Researchers in psychology have extensively studied the relationship between the expression of happiness and depression. Studies indicate that life satisfaction and happiness are negatively associated with depression [59, 60]. However, the lives of people diagnosed with depression are not always depressing; they also experience happiness Bergsma et al. [61]. Building on this research, we aim to identify the expression of happiness in written text, from individuals mentioning a depression diagnosis online to gain insights into the experiences that contribute to their happiness. We address this question in Chapter 6, by conducting a large-scale study examining happy moments from social media posts made by individuals sharing their depression diagnosis. We define a happy moment as the natural language expression of an

experience that evokes feelings of pleasure or self-satisfaction. We develop a comprehensive deep learning framework to extract happy moments from text and annotate them with semantic topics, gender labels, and measures of agency and sociality [9, 19].

## 1.2 Thesis Contributions

In this thesis, we present and expand upon our published research on using computational approaches to analyze mental health through social media content. Our contributions progress from the study of the language of individuals with mental disorders to developing computational methods for detecting mental health problems in texts, images, and videos. Finally, we present our research on analyzing happy moments in social media content. The main contributions of this thesis are:

- A comprehensive study of part-of-speech use in social media data, building on previous psychology and computational linguistics research. Our findings confirm that individuals with self-mentions of depression diagnoses tend to use more pronouns (particularly first-person singular pronouns) and verbs, while using fewer common and proper nouns in their Reddit posts. This contribution was published in our paper:
  - **A.-M. Bucur**, I. R. Podina and L. P. Dinu, *A Psychologically Informed Part-of-Speech Analysis of Depression in Social Media*, Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP) (2021) pp. 199–207
- We are the first to apply offensive language identification techniques to posts made by individuals with depression diagnosis, in order to analyze the use of profanity and offensive language. Our findings indicate that users with self-reported depression diagnoses are more likely to use offensive language in their posts compared to a control group. However, individuals with depression use less targeted offensive language, using instead general profanity. This contribution has been published as:
  - **A.-M. Bucur**, M. Zampieri and L. P. Dinu, *An Exploratory Analysis of the Relation Between Offensive Language and Mental Health*, Findings of the Association for Computational Linguistics: ACL-IJCNLP (2021) pp. 3600–3606
- We participate in the Early Risk Prediction on the Internet Lab, where we contribute to the development of systems for the early detection of risks related to depression, self-harm, and pathological gambling, achieving competitive results compared to other participants. Through our ongoing involvement in the eRisk Lab over several years, we have demonstrated that NLP models can effectively identify early signs of mental disorders by analyzing social

media content. From our participation in the eRisk Lab, we have published four papers:

- **A.-M. Bucur**, A. Cosma and L. P. Dinu, *Early Risk Detection of Pathological Gambling, Self-Harm and Depression Using BERT*, Working Notes of CLEF 2021, Early Risk Prediction on the Internet Lab (eRisk), CEUR 2936 (2021) pp. 938–949
- **A.-M. Bucur**, A. Cosma, L. P. Dinu and P. Rosso, *An End-to-End Set Transformer for User-Level Classification of Depression and Gambling Disorder*, Working Notes of CLEF 2022, Early Risk Prediction on the Internet Lab (eRisk), CEUR 3180 (2022) pp. 851–863
- **A.-M. Bucur**, *Utilizing ChatGPT Generated Data to Retrieve Depression Symptoms From Social Media*, Working Notes of CLEF 2023, Early Risk Prediction on the Internet Lab (eRisk), CEUR 3497 (2023) pp. 662–671
- **A.-M. Bucur**, *Leveraging LLM-Generated Data for Detecting Depression Symptoms on Social Media*, Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2024. Lecture Notes in Computer Science 14958 (2024) pp. 193–204

- We achieve state-of-the-art results in detecting depression from multimodal data, which includes social media text and images. Our approach consists of a flexible time-enhanced multimodal transformer architecture, and we evaluate its effectiveness on two datasets from Twitter and Reddit. Furthermore, our model is interpretable, allowing us to identify the most informative posts from a user's social media timeline. This contribution was published as:
  - **A.-M. Bucur**, A. Cosma, P. Rosso and L. P. Dinu, *It's Just a Matter of Time: Detecting Depression With Time-Enriched Multimodal Transformers*, Advances in Information Retrieval. ECIR 2023. Lecture Notes in Computer Science 13980 (2023) pp. 200–215

- We obtain state-of-the-art results in detecting depression through non-verbal cues from video content by building a multimodal temporal model. By integrating high-level modalities such as facial expression, body language, eye gaze, and voice, our approach significantly enhances the accuracy of depression detection in both in-the-wild contexts and controlled clinical environments. This contribution is based on our published work:
  - D. Gimeno-Gómez, **A.-M. Bucur**, A. Cosma, C.-D. Martínez-Hinarejos and P. Rosso, *Reading Between the Frames: Multi-Modal Depression Detection in Videos From Non-Verbal Cues*, Advances in Information Retrieval. ECIR 2024. Lecture Notes in Computer Science 14608 (2024) pp. 191–209

- We conduct the first large-scale analysis of how individuals with a depression diagnosis express happy moments on social media. Our analysis involves developing a method to detect happy moments in social media texts, creating models for inferring agency and sociality labels, conducting topic modeling, and inferring gender. We discovered a common pattern among both male and female social media users in the depression group: their expressions of happiness were significantly related to topics surrounding social relationships, including friendships, romantic relationships, professional growth, personal memories, and self-disclosure. In addition, we found that both female and male users in the depression group demonstrated a higher level of sociality in their happy moments compared to those in the control group. This contribution is based on two published papers:
  - **A.-M. Bucur**, A. Cosma and L. P. Dinu, *Life Is Not Always Depressing: Exploring the Happy Moments of People Diagnosed With Depression*, Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC) (2022) pp. 4186–4192
  - **A.-M. Bucur**, B. Chulvi, A. Cosma and P. Rosso, *The Expression of Happiness in Social Media of Individuals Reporting Depression*, IEEE Transactions on Affective Computing 16.1 (2025) pp. 360–375

## 1.3 Thesis Outline

The thesis is organized as follows: Chapter 2 provides an overview of the evolution of research on mental health expressions, from analyzing language patterns in clinical interviews and short essays to progressing to computational approaches that analyze language in social media data. We present the data collections used in this thesis and other available resources for research on mental health. In Chapter 3, we explore the manifestations of depression on social media by examining part-of-speech patterns. Additionally, we explore the use of offensive language in social media data of individuals with self-mentions of depression diagnoses. Chapter 4 presents our participation in the Early Risk Prediction on the Internet Lab over three years, during which we addressed the detection of depression, self-harm, and pathological gambling. In Chapter 5, we present the methods developed for detecting depression from multimodal content, including texts, images, and video. Chapter 6 explores happy moments shared on social media by individuals reporting a depression diagnosis. Finally, we conclude our thesis in Chapter 7 and discuss potential future directions.

CHAPTER 2

# Background and Related Work

In this chapter, we provide an overview of expressions related to mental health in written language as well as in multimodal data, including images and videos. We discuss various methods for detecting mental disorders using social media data, beginning with early classical approaches and concluding with recent advancements in Large Language Models (LLMs). Additionally, we highlight how computational models have been used during the pandemic to monitor social media streams to estimate the prevalence of depression in the population. Finally, we present the mental health datasets used in the experiments from this thesis.

## 2.1 Mental Health Expression in Language

Language is a reflection of individuals' mental states. Researchers have been studying the manifestations of mental health problems in language collected from various sources, such as clinical interviews [39], essays [37], suicide notes [62], and social media [40]. For instance, the tendency for self-focused attention, often seen in anxiety and depression, is evidenced through the use of the first-person singular pronoun "I" [37]. Additionally, the cognitive bias towards negative information is reflected through the increased frequency of negative words [38]. Anhedonia, characterized by a lack of interest in daily activities and an inability to feel pleasure, is reflected in a lower frequency of verbs in the future tense [3]. Furthermore, rumination—the repetitive focus on the past, which can lead to emotional distress—is common in several mental disorders, such as depression, anxiety, and post-traumatic stress disorder [63]. This tendency may manifest in language through the frequent use of verbs in the past tense [47]. Based on insights from psychological research, computational methods have been created to automatically identify individuals at risk of mental disorders based on these language patterns. We expand on these approaches in the following sections.

## 2.2 Mental Health Expression in Multimodal Data

Research on the manifestations of mental disorders has primarily focused on language used in social media. However, researchers have also started to explore multimodal information, such as images and videos. For instance, Reece and Danforth [54] compared the images posted online by individuals with depression with the ones from individuals without depression. The authors found that individuals in the depression group tended to post more images featuring faces, but with fewer faces visible in each image. This could suggest an increased self-focus among those with depression and a decrease in social interaction [54, 55]. Additionally, images from the depression group were more likely to be perceived as sad and were characterized by darker, bluer, and grayer tones compared to those from the control group. Guntuku et al. [55] and Uban et al. [64] revealed that users diagnosed with depression also tend to post more images of animals.

Depression, as outlined by the Diagnostic and Statistical Manual of Mental Disorders-V (DSM-V) criteria [30], is characterized by various symptoms, including persistent depressed mood, sleep disturbance, changes in appetite, and a significant loss of interest in once enjoyable activities. A key aspect of depression involves psychomotor changes, which can appear as either agitation or slowing down, and these changes are often linked with other symptoms such as fatigue, low energy, and difficulties with concentration [65]. Non-verbal facial cues related to depression typically include reduced expressiveness [56] and fewer smiles [66]. Regarding body language, individuals with depression may display a downward head tilt [58], longer walking cycles, and a slower pace [67]. Regarding hand movements, people with depression often use fewer gestures [68] and may have impaired gesture performance [69]. Speech patterns among those with depression are marked by a slower speech rate, with longer pauses between words [57], and a noticeable decrease in volume [70]. Eye behavior may also change, evidenced by an increased blinking rate [71], less eye contact, and increased gaze angle [72]. Studying psychomotor changes in depression is essential, as these can help predict how patients will respond to medication [57, 73].

## 2.3 On the State of NLP Approaches to Mental Disorders Detection in Social Media

Modeling mental disorders using social media data is an important topic in NLP research. Depression is one of the most frequently studied mental health issues, with many papers published annually. In Figure 2.1, we show the number of papers on depression detection from social media data published each year. Since the beginning of the COVID-19 pandemic in 2020, there has been an increase in the number of papers on this topic. In total, 409 papers were published between 2017 and 2024. We chose 2017 as the starting year because it was when the first widely-used benchmark datasets were made available (e.g., Yates et al. [74],

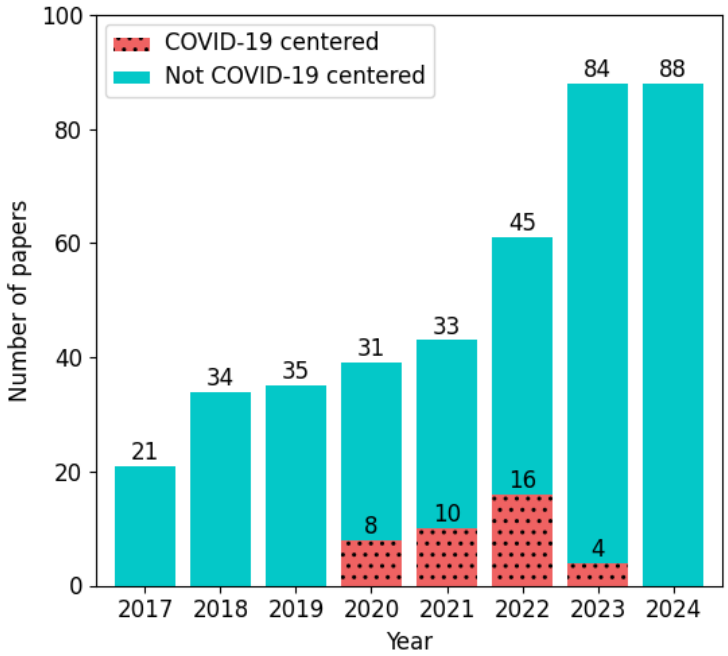



Figure 2.1: Number of papers on depression modeling published each year in peer-reviewed conferences or journals.

Benton et al. [75], Shen et al. [76]) and when shared tasks (e.g., eRisk [77]) have been organized. It was also when the first comprehensive survey on the topic was published [78].

### 2.3.1 The Task of Automatic Mental Disorders Detection

Mental disorder detection can be modeled through various tasks. Following, we make an overview of the most common tasks using social media data (Figure 2.2).
**Binary classification of mental health status** The most common task is modeling mental disorders detection as a binary classification task, where the goal is to predict the presence or the absence of a mental health problem. The binary task can be performed at the post or user levels. In post-level detection, the aim is to predict an individual's mental health status based on a single post. This approach is primarily used to identify suicide ideation [79, 80] and depression [81, 82]. However, relying on only one post for decision-making can often lead to inaccurate predictions. As a result, researchers have shifted to user-level modeling, which uses a larger sequence of posts from an individual's social media timeline. This approach has been successfully used for predicting if an individual is at risk of depression [74, 83, 84], anxiety [85], bipolar disorder [86], or post-traumatic stress disorder [87].

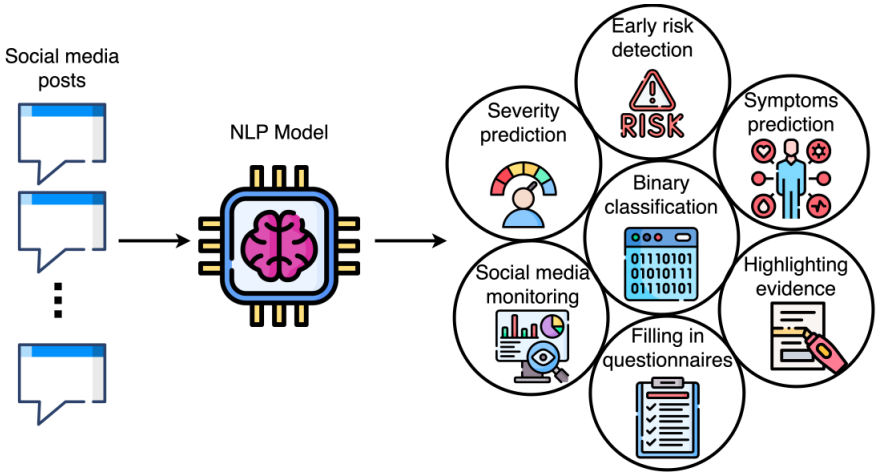



Figure 2.2: Overview of tasks related to detecting mental health problems from social media.

**Early risk detection of mental disorders** Binary classification of mental disorders can be modeled as an early risk prediction task, a method extensively explored in the eRisk Lab [88, 89]. This classification aims to label accurately users as soon as possible. Typically, social media posts are processed sequentially, in chronological order, and a decision is made at each time step. The decision is based on the posts seen so far, allowing the system to either make a prediction or wait for more data before deciding.

**Mental disorders severity prediction** The severity of mental disorders can be approached either as an ordinal regression/classification task or a multiclass classification task. For instance, in the case of depression, the objective is to predict the severity of a social media post as minimal, mild, moderate, or severe [90–92]. Another application is predicting suicide risk, which can be classified into categories such as no risk, low, moderate, or severe risk [93]. Additionally, Gaur et al. [94] model the severity of suicide risk based on the Columbia-Suicide Severity Rating Scale [95]. This scale classifies suicide-related language in a range from supportive language to indications of an actual suicide attempt.

**Mental disorders symptoms prediction** To develop models that are useful for mental health professionals, researchers have turned to exploring social media for symptoms of mental disorders [96–100]. The identified symptoms are based on the Diagnostic and Statistic Manual of Mental Disorders [30] or validated mental health questionnaires, such as PHQ-9 [101] or BDI-II [102]. These models are preferred because they are interpretable and make it easier to explain the predictions.

**Highlighting evidence of mental disorders** Another step towards the explainability of model predictions is identifying evidence for mental disorders. For example, this task can be used to signal the evidence of the suicide risk for a given social media user [15, 22, 103].

**Modeling moments of change** Social media posts can be modeled longitudinally

to detect moments of change in the mental health status of individuals. Previous research has shown that there are specific shifts toward suicide ideation that can be quantified from social media posts [104]. Detecting these moments of change can serve as a warning signal for potential suicidal behavior [6, 105]. Changes in mental health can be indicated by shifts from positive to negative moods or vice versa, as well as by escalations, which refer to a gradual progression of mood [106].

**Filling in validated questionnaires** Mental health indicators from the social media timeline of an individual can be used to fill in validated depression questionnaires. By analyzing the social media data of a user, the task is to estimate various symptoms of mental disorders that are usually assessed through survey-based methods. For instance, the eRisk Lab has proposed using social media data to automatically fill out the Beck's Depression Inventory for depression assessment and the Eating Disorder Examination Questionnaire for evaluating eating disorders risk [89].

**Mental health surveillance** Mental health surveillance involves using social media data to estimate the prevalence of mental disorders in the population. The task uses NLP models to predict the risk of mental disorders on individual social media posts, which are aggregated to estimate the overall prevalence of mental disorders in a population. This approach was primarily used during the COVID-19 pandemic to assess the mental health burden and associated symptoms and has shown results comparable to traditional survey-based methods [107].

### 2.3.2 Shared Tasks for Mental Disorders Identification

The field of mental disorders detection from social media data using NLP approaches has gained increasing interest over time, supported by the organization of workshops and shared tasks specifically focused on mental disorders, such as:

**eRisk** The Early Risk Prediction on the Internet Lab [108] began in 2017 with a pilot task focused on depression detection from textual social media data. Since then, the lab has organized several shared tasks each year. Additionally, the eRisk Lab has expanded its scope to other mental health issues, such as self-harm, eating disorders, or pathological gambling.

**CLPsych** The Workshop on Computational Linguistics and Clinical Psychology[1] started in 2014 and since then has organized several shared tasks focused on issues such as depression and PTSD [87], suicide risk [93], and predicting current and future psychological health based on childhood essays [109].

**LT-EDI** The Workshop on Language Technology for Equality, Diversity, and Inclusion[2] [110, 111] has also organized two shared tasks focused on detecting depression signs from social media data in 2022 and 2023.

**MentalRiskES** The IberLEF shared evaluation campaign hosted the Early Detection of Mental Disorders Risk in Spanish[3] task [112]. This was the first initiative to

[1] https://clpsych.org/

[2] https://sites.google.com/view/lt-edi-2023

[3] https://sites.google.com/view/mentalriskes

propose a shared task on detecting mental disorders in languages other than English. MentalRiskES hosted three different tasks on detecting eating disorders, depression, and anxiety in the Spanish language.

These workshops and shared tasks have been valuable in providing data for research and encouraging the development of novel approaches for detecting mental disorders.

### 2.3.3 Early Approaches

As mentioned earlier, detecting mental disorders through social media data has primarily been approached as a binary classification task. In this context, prior research has utilized various traditional machine learning algorithms, including Logistic Regression [41, 113], Support Vector Machines [84, 114], Naive Bayes [76, 115, 116], Random Forest [3, 48], and Decision Trees [117, 118]. These machine-learning models often rely on hand-crafted features derived from language markers identified in psychological research. Such features may include the use of pronouns [3], the presence of negative or offensive terms [4, 119], and emotional patterns [118, 120]. Additionally, machine learning models rely on psychologically meaningful features, such as Linguistic Inquiry and Word Count (LIWC) categories [121]. LIWC provides a list of dictionary words for more than seventy categories: part-of-speech (e.g., personal pronouns, first-person personal pronouns, nouns, present/future/past verbs, adjectives), psychological processes (e.g., social, affective, cognitive processes), personal concerns (e.g., work, money, religion, death), etc. It is used to detect cues of depression [41, 122, 123], neuroticism [124], or to explore the language of suicidal ideation [125].

### 2.3.4 State-of-the-art

#### Deep Learning

Advancements in technology and the availability of large-scale datasets have led to the increased use of deep neural network architectures to detect mental disorders [126]. Various models, including Convolutional Neural Networks (CNNs) [74, 127], Recurrent Neural Networks (RNNs) [128–130], Long Short-Term Memory (LSTM) networks [131], and Multi-Layer Perceptrons (MLPs) [132], have been employed to identify a wide range of mental disorders. In recent years, pre-trained transformer models have gained popularity for detecting mental health conditions. Models such as BERT [2, 11, 80, 133], XLNet [80, 134], and RoBERTa [7, 50] are among the most commonly used in this field.

Alongside these pre-trained transformer models, models developed for specific domains have also become more popular due to their effectiveness and state-of-the-art performance on downstream tasks. MentalBERT and MentalRoBERTa [135], specialized variants of the BERT and RoBERTa models, have demonstrated promising results in identifying language patterns associated with mental health

issues. Additionally, MentalLongformer and MentalXLNet [136] were introduced to capture the extended context of discussions related to mental health. The PHS-BERT model [137], which was trained on health-related tweets, is used for public health surveillance of social media data. In addition, DisorBERT [138] has been developed to detect signs of mental disorders—such as depression, self-harm, and anxiety—in social media content.

### Large Language Models

Recent advancements in large language models have opened up new potential applications for depression detection. These applications include the assessment of symptoms related to mental disorders [16, 24, 96–100] and the generation of synthetic data [10]. Additionally, LLMs demonstrate significant potential for providing explainable depression detection, as they can generate explanations that mental health professionals can use to interpret the model's predictions [97, 139, 140]. Xu et al. [96] evaluated several LLMs, including Alpaca [141], FLAN-T5 [142], GPT-3.5 [143], and GPT-4 [144], across various tasks related to the identification of mental disorders. In zero-shot and few-shot learning settings, these models showed limited performance. However, when instruction fine-tuning was used, the performance on all tasks improved simultaneously. Moreover, Souto et al. [139] demonstrated that fine-tuned pre-trained language models (such as BERT and MentalBERT) outperform LLMs like GPT-3.5 [143] and Vicuna [145] in depression classification tasks. Yang et al. [97] compared ChatGPT[4] with three supervised baselines and showed that, even if ChatGPT can achieve good results in a zero-shot learning setting, it lacks behind transformer-based specialized models for downstream tasks such as suicide and depression identification from social media data. Amin et al. [99] performed an interpretable mental health analysis through emotional reasoning using ChatGPT on 11 datasets across five tasks related to depression, stress, and suicide ideation. Their results showed that zero-shot ChatGPT performed better than traditional neural network architectures but could not surpass the performance of specialized transformer-based models. To address the limitations of general-purpose LLMs, Yang et al. [140] proposed MentaLLaMA, an open-source instruction-tuned model specifically designed for mental health analysis, which has shown great performance compared to general-purpose LLMs.

Besides mental health assessment, other applications of LLMs are generating and augmenting data [146–148]. Meyer et al. [149] evaluated the synthetic data generated by GPT-3 [150] for conversational tasks. The authors showed that the performance of classifiers trained on synthetic data performed worse than classifiers trained on fewer samples of real user-generated data. The data generated by GPT-3 had less variability than the real data. However, generating synthetic data may be a suitable approach in scenarios with few data or resources.

[4] https://openai.com/blog/chatgpt

### Reliance on Clinical Questionnaires

Regarding the different ways of approaching mental disorders detection, there has been a shift from binary classification models for predicting depression to explainable algorithms that evaluate depressive symptoms using social media data. Researchers use social media content annotated for symptoms identified through psychological assessment tools such as the Patient Health Questionnaire-9 (PHQ-9) and Beck's Depression Inventory-II (BDI-II), or described in the DSM-V, to create methods for detecting mental disorders in an explainable manner. For instance, Zhang et al. [151] developed an automatic method for detecting mental disorders, identifying the specific mental health symptoms contributing to each condition. Moreover, automated depression detection models trained to recognize symptoms of mental disorders have demonstrated greater generalizability compared to traditional binary classification models [152, 153]. To interpret the predictions made by these automated models, some studies use techniques such as attention weights [154, 155], SHAP [2, 156], or LIME [157, 158].

Recently, there have been efforts to augment mental disorder detection methods with information from clinical questionnaires such as PHQ-9 and BDI-II. Nguyen et al. [152] proposed several approaches for depression detection that were constrained by the presence of the symptoms from PHQ-9. The proposed models consisted of a questionnaire model that predicted the PHQ-9 symptoms and a depression model that used the symptom features for prediction. The authors showed that the models constrained on PHQ-9 had comparable performance to unconstrained methods, could better generalize to other datasets, and are interpretable. Similarly, Zhang et al. [151] performed symptom-assisted mental disorders identification, achieving better results than baselines that use only text. Furthermore, their method was interpretable and provided symptom-based explanations for several mental health disorders, such as depression, anxiety, bipolar disorder, obsessive-compulsive disorder, eating disorders, ADHD, and post-traumatic stress disorder. Psychiatric scales were used to screen individuals at risk using Hierarchical Attention Networks for the early detection of depression [159]. Liu et al. [160] crawled data from different subreddits corresponding to 13 depression symptoms (e.g., *r/insomnia*, *sleep* for sleep problems, *r/chronicfatigue*, *r/fatigue* for fatigue, etc.). Different models were trained on the data to detect each symptom. The predictions of these symptom detection models on Facebook data were validated against PHQ-9, General Anxiety Disorder-7 (GAD-7), and UCLA Loneliness Scale (UCLA-3) filled in by individuals. The authors showed that the automatically predicted symptoms were significantly associated with the symptoms checked by the self-report surveys, except for fatigue.

### Multimodal Approaches

Even if deep learning methods such as CNNs [74, 161], LSTMs [46, 162] and transformer-based architectures [2, 49, 50] achieved good results on depression

detection using only the textual information from users' posts, multimodal methods incorporating visual features achieve even greater results [51–53]. Shen et al. [76] collected the first user-level multimodal dataset from social media for identifying depression with textual, behavioral, and visual information from Twitter users and proposed a multimodal dictionary learning method. The same Twitter dataset was later explored by Gui et al. [163], who used a cooperative multi-agent reinforcement learning method with two agents for selecting only the relevant textual and visual information for classification. An et al. [164] proposed a multimodal topic-enriched auxiliary learning approach, in which auxiliary tasks improve the performance of the primary task on depression detection on visual and textual topic modeling. By not considering the time component, the above methods assume that social media posts are synchronous and sampled at regular intervals [13]. Realistically, posts from online platforms are asynchronous, and previous studies have shown differences in social media activity, partly due to the worsening of depression symptoms at night [40, 165, 166]. Motivated by this, methods that use time-adapted weights [128] or Time-Aware LSTMs [167, 168] to include the time component of data for mental health problems detection report higher performance.

Depression detection from video involves automatically extracting temporal and spatial features and processing them with a classifier [14, 169]. Some low-level features extracted using classical image and audio processing algorithms are loudness, spectral flux, phoneme duration, pitch slopes for audio and facial landmarks, facial action units [170], and histogram of oriented gradients features for video content [171, 172]. High-level semantic features include facial expressions, smile intensity and duration, head movement, etc. [169]. Most current methods proposed for depression detection [173–176] use facial landmarks and low-level acoustic descriptors [177], disregarding the temporal information [174, 177, 178]. More recently, multimodal large language models offer a promising paradigm for video classification, with approaches such as MiniGPT-4 [179] and PandaGPT [180] obtaining good results.

### 2.3.5 Datasets for Depression Detection

The availability of data collections has advanced the development of state-of-the-art depression prediction models. As depression is the most prevalent mental disorder worldwide, most data collections focus on depression. In this section, we present the particularities of the datasets for depression detection published between 2017-2024.[5] Out of the 409 papers published during this period, 81 of them collect and annotate new data from online platforms. In Figure 2.3, we present the most used datasets in the 409 papers. In these papers published between 2017 and 2024, most of the works have used the datasets from the eRisk Lab [77, 88, 181–184], the LT-EDI Workshop (DepSign dataset [92]), or the CLPsych 2015 Shared Task

[5] For the publicly available datasets, links to the corresponding resources can be found at: https://github.com/bucuram/depression-datasets-nlp

dataset [87]. All the aforementioned datasets were released as part of shared tasks or competitions, and the data was a valuable resource that was further used after the end of the shared task. Other benchmark datasets are from Shen et al. [76], Pirina and Çöltekin [81], RSDD [74], or SMHD [113].

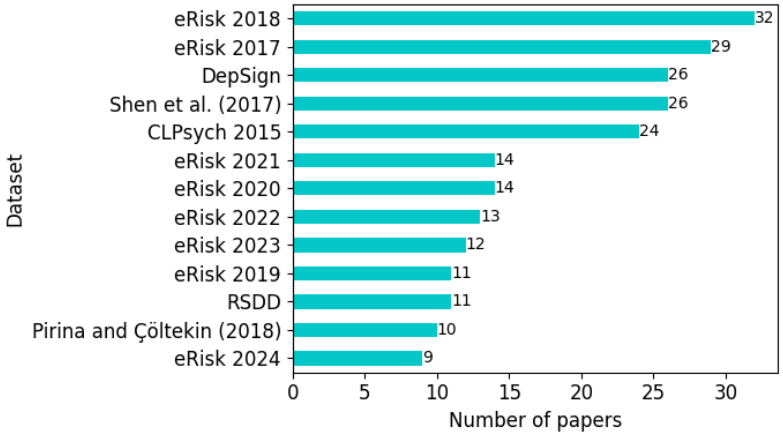



Figure 2.3: The most used datasets for depression detection.

We present the particularities of the datasets used for depression detection in Table 2.1. The table includes information about the platform used for data collection, the annotation procedure and the level of annotation, the labels provided for the data, the size of the dataset, and its availability. The table was included in our overview paper (Bucur et al. [21]).

**Availability** Given the sensitive information in the datasets for depression detection, their availability varies. The labels for data availability were inspired by the work of Harrigian et al. [185] and are explained below:

`FREE` - The dataset is publicly available and hosted online for anyone to access.

`AUTH` - The data can be accessed by contacting the paper's authors.

`DUA` - The data is available only after a data usage agreement is signed. Sometimes, authorization from an Institutional Review Board (IRB) may be needed.

`UNK` - The dataset's availability is unknown; the authors do not mention if the data is available to the research community.

`N/AV` - The dataset is no longer available or cannot be shared due to ethical considerations.

Out of the 81 papers proposing new datasets, 18 are publicly available and hosted online for anyone to use, 17 can be made available after signing a data usage agreement, and 13 collections can be made available by contacting the authors of the dataset. The availability of the rest of the datasets is unknown, and four of them are not available.

**Platform** In Figure 2.4, we present the top five social media platforms used for gathering datasets for depression detection. Twitter and Reddit are the most

used online platforms for collecting data due to the availability and easy access to dedicated APIs. However, due to relatively recent changes in terms of service and API rate limits for Twitter and Reddit, data acquisition on these online platforms was made more difficult. Moreover, these recent changes might hinder the reproduction of datasets where the authors make only the Twitter or Reddit IDs available and not the raw texts. Furthermore, these changes make collecting new data challenging and time-consuming.

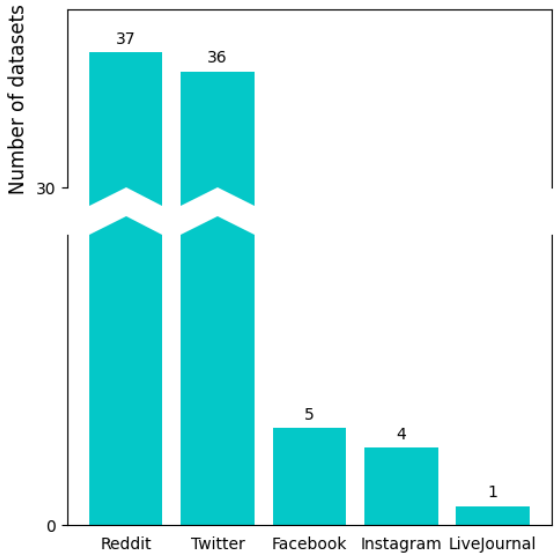



Figure 2.4: The most used platforms for collecting mental-health related data.

**Annotation procedure and labels** For depression detection from social media data, the most common method for annotation is to automatically label users binary by the presence or absence of mentions of depression diagnosis in their posts. Using this approach, large datasets with hundreds of thousands of users can be collected. However, even if annotators manually inspect the posts with mental health disclosures, there is no way to verify whether the disclosure is authentic and whether the user is telling the truth. Furthermore, regarding the control group, in which users do not have any mentions of a mental health diagnosis, their actual mental health status is unknown; we cannot imply that they do not have any mental disorders only by the absence of self-disclosure. It is important to note that relying on self-reported diagnoses to collect data about mental health can lead to self-selection bias [45]. This means that the collected data may only represent individuals who openly discuss their mental health diagnosis and may not accurately reflect the entire population of people diagnosed with mental disorders. From the datasets in Table 2.1, 33 rely on the self-disclosure of diagnosis for the annotation procedure.

Another method for annotating users is asking them to fill in validated self-report questionnaires, such as the BDI-II or PHQ-9. However, even if psychometric tools

produce a more reliable assessment of depression, fewer people are willing to participate in the data collection, resulting in small sample sizes. Only nine datasets in Table 2.1 rely on self-report questionnaires for the annotation procedure.

Data collection can also rely on annotations made manually by mental health experts or students. Unlike annotations derived from self-report questionnaires or self-mentions, which are made at the user level, most procedures for manual annotations are performed at the post level. Manual annotation is used to label the data binary (depression vs. control), to label data for depression severity (no signs of depression, mild, moderate, severe, etc.), and for symptoms measured by different validated questionnaires such as PHQ-9 and BDI-II, or symptoms from the DSM-V. Recently, datasets have shifted from binary labeling to depression symptom-based labeling, leading to the development of explainable methods for depression detection. From the datasets in Table 2.1, only 20 of them rely on manual annotations.

Table 2.1: List of available datasets for depression detection using data posted on online platforms. *Denotes that the dataset contains data collected during the COVID-19 pandemic.

| Dataset | Platform | Level | Annotation Procedure | Label | Dataset Size | Availability |
|---|---|---|---|---|---|---|
| Multitask [75] | Twitter | USER | Self-disclosure | Labels for multiple disorders | 9.5K users | UNK |
| RSDD [74] | Reddit | USER | Self-disclosure | Binary | 116K users | N/AV |
| Aldarwish and Ahmad [186] | Twitter, Facebook, LiveJournal | POST | Manual annotation | Binary, DSM-IV symptoms | 6.7K posts | UNK |
| Reece and Danforth [54] | Instagram | USER | CES-D | Binary | 166 users | UNK |
| Shen et al. [76] | Twitter | USER | Self-disclosure | Binary | 2.8K users | FREE |
| 160Users [114] | Twitter | USER, POST | Self-disclosure | Binary | 160 users, 8K posts | AUTH |
| SAD corpus [187] | Twitter | POST | Manual annotation | Symptoms, psychological stressors | 9.3k posts | UNK |
| Vedula and Parthasarathy [118] | Twitter | USER | Depression-related keywords | Binary | 150 users | UNK |
| eRisk 2017 [77] | Reddit | USER | Self-disclosure | Binary | 887 users | DUA |
| Yazdavar et al. [188] | Twitter | USER | Self-disclosure | Binary | 47K users | UNK |
| Rojas-Barahona et al. [189] | Koko Platform | POST | Manual annotation | CBT Concepts | 4035 posts | AUTH |
| Pirina and Çöltekin [81] | Reddit | POST | Subreddit participation | Binary | 3.6K posts | FREE |
| Eichstaedt et al. [41] | Facebook | USER | Medical records diagnosis | Binary | 683 users | UNK |
| Seabrook et al. [190] | Twitter, Facebook | USER | PHQ-9 | Depression severity | 78 users | N/AV |
| Ricard et al. [191] | Instagram | USER | PHQ-8 | Binary | 749 users | UNK |
| TRT [83] | Reddit | USER | Self-disclosure | Binary | 12K users | UNK |
| eRisk 2018 [181] | Reddit | USER | Self-disclosure | Binary | 1.1K users | DUA |
| Loveys et al. [122] | 7 Cups of Tea | USER | Self-disclosure | Binary | 1.9K users | UNK |
| Chen et al. [192] | Twitter | USER | Self-disclosure | Labels for multiple disorders | 7.9K users | UNK |
| Chen et al. [193] | Twitter | USER | Self-disclosure | Binary | 7K users | UNK |
| RSDD-Time [194] | Reddit | USER | Self-disclosure | Labels for multiple disorders | 598 users | N/AV |
| Islam et al. [195] | Facebook | POST | - | Binary | 7K posts | FREE |
| SMHD [113] | Reddit | USER | Self-disclosure | Labels for multiple disorders | 350K users | N/AV |

| Dataset | Platform | Level | Annotation Procedure | Label | Dataset Size | Availability |
|---|---|---|---|---|---|---|
| Gui et al. [163] | Twitter | USER | Self-disclosure | Binary | 2.8K users | UNK |
| Guntuku et al. [55] | Twitter | USER | BDI | Binary | 887 users | UNK |
| Almouzini et al. [196] | Twitter | USER, POST | Manual annotation | Binary | 89 users | UNK |
| eRisk 2019 [182] | Reddit | USER | BDI-II | BDI filled-in | 20 users | DUA |
| Owen et al. [82] | Twitter | POST | Manual annotation | Binary | 1K posts | FREE |
| Bathina et al. [197] | Twitter | USER | Self-disclosure | Binary | 1.2K users | AUTH |
| Ríssola et al. [198] | Reddit | POST | Self-disclosure, heuristics | Binary | 14K posts | DUA |
| Birnbaum et al. [199] | Facebook | USER | Medical records diagnosis | Binary | 223 users | AUTH |
| D2S [200] | Twitter | POST | PHQ-9 | PHQ-9 symptoms | 12K posts | AUTH |
| eRisk 2020 [183] | Reddit | USER | BDI-II | BDI filled-in | 90 users | DUA |
| Tabak and Purver [201]* | Twitter | USER | Self-disclosure | Binary | 5K users | UNK |
| Yazdavar et al. [84] | Twitter | USER | Manual annotation | Binary | 8.7K users | DUA |
| Haque et al. [80] | Reddit | POST | Subreddit participation | Depression vs. suicide | 1.8K posts | FREE |
| Chiu et al. [128] | Instagram | USER | Depression-related keywords | Binary | 520 users | UNK |
| Arachchige et al. [132] | Online forums | POST | Manual annotation | Depression severity | 2.1K posts | UNK |
| Sherman et al. [202] | Reddit | USER | Self-disclosure | Binary | 31K users | DUA |
| eRisk 2021 [88] | Reddit | USER | BDI-II | BDI filled-in | 170 users | DUA |
| Pirayesh et al. [203] | Twitter | USER | Self-disclosure | Binary | 817 users | AUTH |
| Guo et al. [204] | Reddit | USER | Self-disclosure | Labels for multiple disorders | 7.9 K users | UNK |
| Zhang et al. [205]* | Twitter | USER | Self-disclosure | Binary | 5K users | UNK |
| Zhou et al. [206]* | Twitter | USER | Self-disclosure | Binary | 1.8M posts | UNK |
| Safa et al. [207] | Twitter | USER | Self-disclosure | Binary | 1.1 K users | AUTH |
| Naseem et al. [90] | Reddit | POST | Manual annotation | Depression severity | 3.5 K posts | FREE |
| PsySym [151] | Reddit | USER, POST | Automatic and manual annotation | DSM-5 symptoms for multiple disorders | 26K users, 8.5K posts | AUTH |
| MHB [208] | Mental health forums | USER | Forum participation | Only depression | 9.3K users | FREE |
| CAMS [209] | Reddit | POST | Manual annotation | Causes for depression | 3.1 K posts | FREE |
| Sotudeh, Goharian and Young [210] | Reddit | POST | Subreddit participation | Summarization | 24 k posts | DUA |

| Dataset | Platform | Level | Annotation Procedure | Label | Dataset Size | Availability |
|---|---|---|---|---|---|---|
| Sampath and Durairaj [92] | Reddit | POST | Manual annotation | Depression severity | 16K posts | FREE |
| eRisk 2022 [184] | Reddit | USER | Self-disclosure | Binary | 3.1K users | DUA |
| Monreale et al. [211] | Reddit | POST | Subreddit participation | Labels for multiple disorders | 16 K posts | UNK |
| PRIMATE [212] | Reddit | POST | Manual annotation | PHQ-9 symptoms | 2K posts | DUA |
| PsycheNet-G [213] | Twitter | USER | Self-disclosure | Binary | 591 users | UNK |
| Twitter-STMHD [214] | Twitter | USER | Self-disclosure, manual annotation | Labels for multiple disorders | 33K users | FREE |
| multiRedditDep [64] | Reddit | USER | Self-disclosure | Binary | 3.7K users | AUTH |
| Davis et al. [215]* | Reddit | USER | Subreddit participation | Binary | 81K users | UNK |
| Fernández-Barrera et al. [216]* | Flickr | POST | Depression tags | Only depression | 14.5K posts | UNK |
| Cha et al. [217]* | Twitter, Everytime | POST | Lexicon-based automatic annotation | Binary | 26M posts, 22K posts | AUTH |
| DEPTWEET [91] | Twitter | POST | Manual annotation | Depression severity | 40K posts | FREE |
| Alavijeh et al. [218] | Twitter | USER | Self-disclosure | Labels for multiple disorders | 1.5K users | FREE |
| Adarsh et al. [158] | Reddit | POST | Subreddit participation | Binary | 60K posts | UNK |
| Liu et al. [160] | Reddit | POST | Subreddit participation | Symptoms | 1.3M posts | FREE |
| BDI-Sen [153] | Reddit | POST | Manual annotation | BDI-II symptoms | 4.9K posts | DUA |
| Song et al. [219] | Reddit | POST | Subreddit participation | Labels for multiple disorders | 85K posts | UNK |
| RedditCE [220] | Reddit | POST | Manual annotation | Emotion-cause labels | 35K posts | FREE |
| Liu et al. [221] | Reddit, Twitter | USER | Self-disclosure | Binary | 205K users, 255 users | UNK |
| RESTORE [222] | Reddit, Twitter, Pinterest | POST | Manual and automatic annotation | PHQ-9 symptoms | 9.8K images | AUTH |
| Zogan et al. [223]* | Twitter | USER | Self-disclosure | Binary | 1.4K users | UNK |
| Wu et al. [224]* | Twitter | USER | Self-disclosure, manual annotation | Binary | 10K users | DUA |
| DepreSym [100] | Reddit | POST | Manual annotation | BDI-II symptoms | 21K posts | DUA |
| Villa-Pérez et al. [225] | Twitter | USER | Self-disclosure | Labels for multiple disorders | 6K users | DUA |
| HelaDepDet [226] | Twitter, Reddit | POST | Manual annotation | Depression severity | 40K posts | FREE |

| Dataset | Platform | Level | Annotation Procedure | Label | Dataset Size | Availability |
|---|---|---|---|---|---|---|
| Anshul et al. [227]* | Twitter | USER | Self-disclosure, Manual annotation | Binary | 1.5K users | FREE |
| Alhamed et al. [228] | Twitter | USER | Manual annotation | Before/After diagnosis | 120 users | FREE |
| Milintsevich et al. [229] | Reddit | POST | Manual annotation | Anhedonia | 167 posts | DUA |
| MentalHelp [16] | Reddit | POST | Automatic annotation | Binary | 14M posts | FREE |
| Lee et al. [230] | Reddit | USER | Manual annotation | Binary | 1K users | DUA |
| Beniwal and Saraswat [231] | Instagram | POST | Manual annotation | Binary | 10K posts | AUTH |
| Tumaliuan et al. [232] | Twitter | USER | PHQ-9 | Binary | 72 users | AUTH |

## 2.4 A Post-COVID-19 Outlook

Starting on March 11th, 2020, the COVID-19 global pandemic brought several negative social and economic consequences with serious repercussions on mental health all over the world. Prolonged social isolation and fear of contagion combined with harsh and unprecedented government measures such as school and business closures have created the conditions for a depression epidemic. World Health Organization (WHO) estimates that the prevalence of depression worldwide has spiked by around 25% in the first year of the pandemic alone.[6] Studies on the prevalence of depression published throughout the pandemic estimate even higher increase rates of depression of above 50% [233, 234]. On May 5th, 2023, after more than three years since the beginning of the pandemic, WHO announced that COVID-19 is no longer "a public health emergency of international concern" [235]. The WHO announcement marked, at least in practice, the end of the COVID-19 pandemic, but the negative effects of the global public health emergency are likely to continue in the years to come.

In this section, we present a post-pandemic outlook on depression detection in social media using NLP, as symptoms of depression were the most extensively studied during the pandemic, based on our published work (Bucur et al. [20]). We present papers published between 2020 and 2023, specifically focusing on studies that conduct mental health-related analyses in the context of the COVID-19 pandemic using social media data. Out of the 231 papers published during the COVID-19 pandemic, only 38 focused on using social media data from that time to assess the mental health burden on the population.

In Figure 2.5, we propose a taxonomy of the 231 papers on depression detection published during the 2020-2023 period. Most articles (n = 193) tackling depression detection from social media data proposed general methods or datasets; only a small number of papers (n = 38) chose to work on assessing the impact of the global COVID-19 pandemic on the population's mental health. In this section, we present the COVID-19-related studies according to the classification from Figure 2.5: studies monitoring social media posts and works proposing methods for depression detection in the context of the pandemic.

### 2.4.1 Monitoring Social Media Streams during COVID-19 Pandemic

Most studies that use different social media platforms to monitor depression manifestations during the pandemic focus on data collected from Twitter. Other social media platforms of interest were Reddit, Flickr, and Facebook. For measuring the impact of the pandemic on mental health, most papers use keyword-based approaches for collecting depression-related content. Few works use machine

[6] https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide

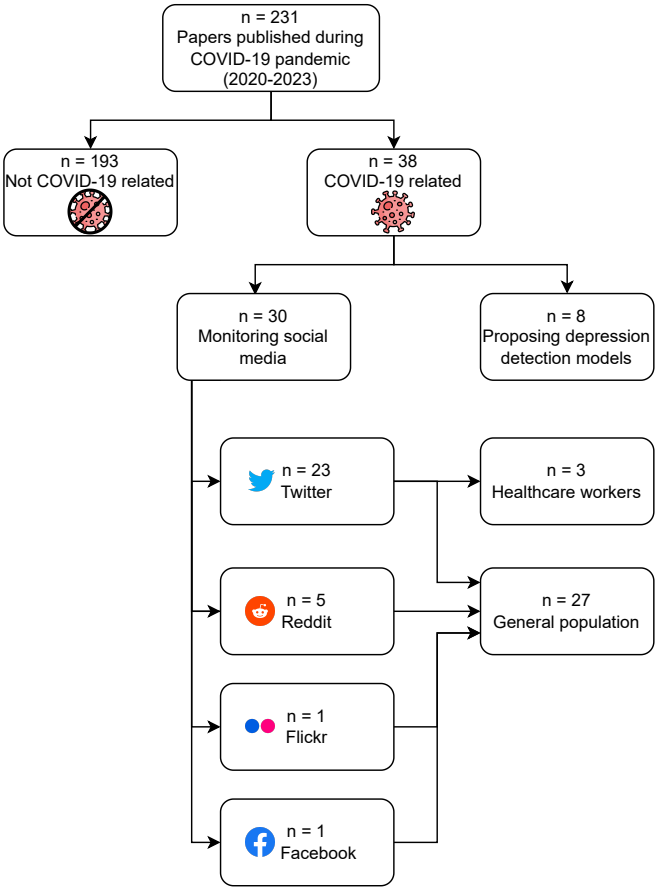



Figure 2.5: Taxonomy of papers published during the pandemic (2020-2023). Only 38 out of the 231 papers addressed depression detection in the context of the pandemic.

learning models to predict user- and post-level depression.

**Twitter** During the COVID-19 pandemic, Twitter was the leading social media platform for mental health surveillance. Two main approaches were chosen for real-time assessment of population-level mental health: keywords-based methods for retrieving and analyzing depression-related content, and machine learning approaches for detecting the presence or absence of mental disorders symptoms. Keyword-based analyses revealed a rise in depression-related tweets at the beginning of the pandemic [236, 237]. Furthermore, using vocabulary-based methods, Cohrdes et al. [107] extracted and monitored depression symptoms from Twitter data, such as depressed mood, sleep problems, fatigue, feelings of worthlessness, and diminished interest. With regard to automatic methods for social media monitoring, the studies published during the pandemic used various methods for estimating population-level depression from classical machine learning models [201, 206, 233, 238, 239] to state-of-the-art deep learning models [205, 234, 240].

**Reddit** Regarding Reddit, Biester et al. [238] analyzed the social media posts from three subreddits, `r/Anxiety`, `r/depression`, and `r/SuicideWatch`, during January 2017 and May 2020, to measure the effect of the pandemic on posting patterns. Unlike previous studies on Twitter, which showed an increase in depression-related posts, the activity on `r/depression` decreased at the beginning of the pandemic. Topic analysis revealed that the posts from home, family, and children topics increased significantly in the `r/depression` subreddit. Similar to previous work, Biester et al. [241] compared the activity in `r/Anxiety`, `r/depression`, and `r/SuicideWatch` subreddits with control subreddits and showed that, while the activity in mental health-related subreddits decreased during the pandemic, the activity in control subreddits increased. Chen et al. [242] also showed that the sentiment in the `r/depression` subreddit became more negative during the pandemic.

As opposed to the previous studies, which monitored specific subreddits, Wolohan [233] monitored random users active in `r/AskRedit` to avoid any biases from users' membership in specific depression-related subreddits. The authors collected a novel dataset with users' Reddit posts from three distinct time spans: the first six months of 2018 and 2019 and the first four months of 2020. An LSTM model using fastText embeddings was trained on the Off-Topic Depression dataset [83] to predict user-level depression in each of the three periods of time analyzed. Compared to the periods from 2018 and 2019, the data showed an approx. 50% increase in depression rate during the first four months of 2020 when the pandemic started. Bajaj et al. [243] monitored Reddit and showed that 6.4% of users who did not have any depression-loaded posts before the pandemic began experiencing depression symptoms as detected by machine learning models during the pandemic.

**Facebook** Sukhwal and Kankanhalli [244] studied posts from Facebook groups to determine the impact of containment policies on the mental well-being of the population. The analysis of topics discussed in these posts revealed that issues related to depression and suicide emerged after the lockdown, along with concerns related to the health and safety of frontline workers. Even after the containment policies were relaxed, the discussion of depression continued to increase, indicating that people were still struggling with the effects of the pandemic on their mental health.

**Flickr** In contrast to prior efforts that studied textual information, Fernández-Barrera et al. [216] analyzed the images with depression tags and their descriptions posted on Flickr. The pre-pandemic data was collected from January 2018 to December 2019, while the post-pandemic data was published between January 2020 and December 2021. The pre-pandemic posts contained tags associated with depression symptoms, such as "sad", "melancholic", and "gloomy", while post-pandemic tags were related to feelings of loneliness and isolation. Furthermore, the visual analysis of the images revealed that there were more prevalent pre-pandemic images in the outdoor context than post-pandemic images with more indoor photos,

which was expected due to the limited mobility imposed on the population.

### 2.4.2 How has the Pandemic Affected the Population?

Studies for real-time mental health assessment in social media were performed to understand the impact of the COVID-19 pandemic on the population. In this section, we explore the research published during the pandemic to understand how healthcare workers and the general population from specific geographical regions were affected by the pandemic. In addition to the general population, healthcare workers were monitored since they were at risk for developing mental health problems due to increased workloads and direct involvement in treating those affected by the virus [245].

#### General Population

Studies were carried out on population mainly from North America (US [205, 234, 239, 246–250], Canada [251]) and South America (Argentina [237]). Fewer studies focused on the European continent [107, 201, 252], Australia [206] or Asia (Japan [253], Singapore [244], Indian subcontinent [254]). Other studies weren't focused on a particular location [216, 224, 233, 236, 238, 240, 241, 243, 255–257]. These studies have shown an increase in the population rate of depression [233] and in the depression symptoms reported by users in social media [107, 236]. Below, we present the findings from the studies that used monitoring methods to assess the impact of the pandemic on the population from the geographical regions mentioned. Even if there was a high prevalence of depression due to the pandemic in the population from African countries, as shown by studies from psychology [258], no computational works were focusing on this region, except for Wang et al. [259], which monitored depression trends globally. Compared to other regions analyzed, Wang et al. [259] found fewer tweets geolocated in Africa and South America.

**North America** In the **US**, the depression evolution was aligned with the national prevalence of COVID-19 cases [246, 248, 250]. Levanti et al. [250] used classical machine learning models to detect depression in tweets from 7 major US cities with different response protocols and lockdown policies (Atlanta, Chicago, Houston, Los Angeles, Miami, New York, and Phoenix) between January to September of both 2019 and 2020. The depression evolution in these cities was aligned with the national prevalence of COVID-19 cases, rather than local trends or stay-at-home orders. Kaseb, Galal and Elreedy [234] showed that depressed tweets were more frequent than non-depressed ones and that the depression level increased from February 2020 to November 2021. Zhang et al. [248] analyzed the mental health-related concerns on Twitter during the pandemic from different demographic groups and showed that Male, White, and aged between 30 and 49 years old users expressed most mental health concerns on Twitter. Moreover, topic analysis using Latent Dirichlet Allocation (LDA) revealed that most concerns shared on social media were

related to social isolation, death toll, politics and policies. Saha et al. [247] used the Sparse Additive Generative Model (SAGE) [260] to identify language indicative of mental disorders and compared the 2020 period to 2019 to monitor for changes in mental health at the beginning of the COVID-19 pandemic in the US. The results showed a 21.32% increase in anxiety, 19.73% in suicidal ideation, 10.18% increase in depression, and 3.76% in stress. The expressions of social support also became more prevalent during the pandemic, with a 4.77% increase in emotional support and 4.78% in informational support. After the initial increase, the temporal analysis indicated a decline in mental health symptoms and support expressions.

Tshimula et al. [251] collected Twitter data during the stay-at-home period in **Canada** using COVID-19 related keywords and #StayAtHome hashtag. LDA-based topic analysis on the Twitter data revealed that users were engaging more in depression-related topics during the stay-at-home period, with a rise in the second week of March, corresponding to the onset of the Canadian lockdown. Towards the end of the lockdown in May, Twitter users showed a decrease in participation in depressed-related topics.

**South America** Tommasel et al. [237] used a lexicon-based approach to analyze Twitter data from **Argentina** between March and August 2020 to monitor the changes in mental health during the lockdown. The rise in depression-related posts began with the confirmation of the first COVID-19 case in Argentina and the establishment of the lockdown. Moreover, a high prevalence of depression-related posts was observed when the lockdown extension was announced, with Twitter users expressing health and economic concerns.

**Europe** Tabak and Purver [201] analyzed the temporal depression dynamics in relation to the lockdowns from **France**, **Germany**, **Italy**, **Spain** and the **United Kingdom**. One Bi-directional LSTM model with self-attention was trained on Twitter data from 2015 to October 2019 for each of the five locations of interest to predict depression. The models were further used to predict depression for each day from December 1, 2019, until May 15, 2020, using all the posts from a given day. The authors compared the depression rate before and during the lockdown for each country. At the beginning of the lockdown, there was an increase in depressed tweets from French, Italian, German, and Spanish users. Regarding the UK, there had been an increase in depression before the national lockdown, which afterward plateaued when the lockdown began. The easing of the lockdown restrictions corresponded to an improvement in mental state for the French and Spanish populations, with the depression rate in Italy and Germany further increasing. In contrast, the UK population was agnostic to the change. Another study on the **UK** population between July 23, 2020, and January 6, 2021, had shown that the amount of depression-related tweets increased before the second national lockdown began, followed by a decrease and another increase with the enforcement of the third national lockdown [252].

In a study by Cohrdes et al. [107], depression symptoms on Twitter were compared to survey-based self-reports using PHQ-8 between January and July 2020

in **Germany**, before, during, and after the first lockdown. Posts with depression symptoms were extracted from Twitter data using a vocabulary-based method. The study found a moderate to high correlation between depression symptoms measured through survey and Twitter data for most of the symptoms of interest, such as depressed mood, sleep problems, fatigue, and feelings of worthlessness, except for diminished interest. Unlike previous studies that showed high depressive symptoms during the pandemic and lockdowns, the German population demonstrated a temporary decrease in some symptoms during the lockdown. The study highlights the potential of Twitter data to monitor users' mental states online and provide early intervention measures.

**Asia** Bashar [253] compared the depression levels of people from **Japan** before and during the COVID-19 pandemic. A logistic regression model was used for classifying tweets into depressive and non-depressive categories. The authors showed a higher prevalence of tweets categorized as depressive during the pandemic, with 50.37% tweets classified as depressive out of the total COVID-19-related tweets. The depression level of Japanese Twitter users increased considerably after the peak of COVID-19 deaths in Japan.

Gour et al. [254] performed real-time depression analysis on Twitter streams in countries from the Indian subcontinent (**India, Bhutan, Pakistan, Nepal**). The authors monitored Twitter users' emotions and sentiments using lexicon and machine learning approaches (Decision Trees, Naive Bayes) to understand the depression dynamics of people from the countries above. Pakistan had the largest number of negative sentiment tweets, while also having high levels of posts with fear and anger.

Sukhwal and Kankanhalli [244] investigated Facebook posts from **Singapore** to measure the influence of containment policies on the population's mental health. Topic analysis showed the emergence of depression and suicide topics after the lockdown was announced in Singapore. With the relaxation of containment policies, the topic of depression became even more prevalent, indicating that people were still suffering due to the pandemic.

**Australia** Zhou et al. [206] studied the depression dynamics in the population of New South Wales, **Australia**, using Twitter data from January to May 2020. A logistic regression model with TF-IDF, emotional, topic, and domain-specific features was used to detect depression in Twitter data posted during the pandemic. The results showed that the community depression level increased proportional to the number of COVID-19 cases and then decreased after the end of March when the COVID-19 cases peaked.

As opposed to previous studies, which monitor tweets geolocated in a specific region, Wang et al. [259] collect and analyze depression and COVID-19-related posts from different geographical areas: North America, South America, Europe, Asia, Africa, and Oceania. The study aimed to uncover the depression trends before and after the release of the COVID-19 vaccine. A keyword-based analysis revealed that different themes were present in the depression-related posts during

the pandemic. Still, after the release of the vaccine, a common theme of returning to normal social activities was prevalent.

#### Healthcare Workers

Healthcare workers were more susceptible than the general population to mental illness during the pandemic, as they were on the frontline taking care of patients during the emergency situation [261]. All studies on the mental well-being of healthcare workers used Twitter data geolocated in the **US**. Fine et al. [239] extracted and analyzed real-time mental health information from the social media platforms of healthcare providers. The authors collected and compared a sample of posts from healthcare professionals to posts from a control group comprised of people not working in the health sector. Logistic regression models were trained on n-gram features to predict if a post was written by a person suffering from depression, anxiety, or at risk of suicide. The study showed that healthcare professionals experienced higher levels of depression, anxiety, and suicide risk than the control group before the lockdown began. During the lockdown, the general population showed higher anxiety levels than healthcare providers. However, healthcare workers showed higher levels of suicide risk at the beginning of the lockdown. The changes in depression levels of both groups did not differ significantly during the period analyzed.

Another study monitored Twitter data from February 2020 to April 2022 to measure the impact of lockdown policies on the population's mental health [249]. The authors compared mental health and COVID-19-related posts of healthcare workers to the ones from the general population. A lexicon-based approach was used to categorize mental health-related posts into four categories: anxiety, depression, addiction, and insomnia. All the mental health problems studied were more prevalent in the posts of healthcare professionals than in the general population.

Agarwal et al. [262] analyzed Twitter posts from emergency physicians and resident physicians in the US before and during the pandemic using different psychological measures for anxiety, depression, loneliness, and anger. The authors showed an increase in posts related to the four psychological measures during the pandemic. Furthermore, topic modeling analysis revealed that, during the pandemic, physicians expressed more concerns regarding pandemic response, healthy behaviors during the pandemic, vaccines, unstable housing, and providing emotional support.

### 2.4.3 Depression Detection in the Context of the COVID-19 Pandemic

Few papers have proposed novel methods for depression detection in the COVID-19 context. We expand on some of the works in this section.

Given the increase in depression levels caused by the pandemic, Wu et al. [224] proposed a method for early detection of depression for COVID-19 patients. One of the contributions of the paper was the DepCOV dataset. The data collection

consisted of retrieving users with COVID-19 infection mentions from Twitter. These users were split into two groups: the control group, which did not mention depression before or after the COVID-19 diagnosis, and the group with depression signals, which had at least three depression-related posts after the infection, but none before the infection. The authors proposed a novel transformer-based method, Mood2Content, to detect depression in COVID-19 patients using content and mood representations as early as possible. The model trained on the DepCOV dataset achieved a performance of 0.93 AUROC in predicting COVID-19 patients at risk of depression two weeks before their first depression mention. Topic analysis on the social media content with depression mentions from COVID-19 patients revealed that concerns related to the pandemic, such as lockdown, government, policies, and treatment, were predominant.

Cha et al. [217] experimented with different architectures (BiLSTM, CNN, BERT-based) for the early detection of depression in English, Korean, and Japanese languages during the pandemic. Twitter data was collected from January 2019 to March 2021 to train the models. For each of the three languages, a lexicon-based approach was used to label posts from users in one of the two categories: depressed and non-depressed. The proposed models were trained on the labeled data to predict depression at the post-level and obtained good results of over 98% F1-score. The authors then tested the models' generalizability by applying them to posts from Everytime, a community platform for Korean undergraduate students, during and before the pandemic. While the performance of the CNN and BiLSTM-based architectures degraded when predicting on the Everytime data, the BERT-based model had the greatest performance for this use case, with an F1-score of 64%.

Zogan et al. [223] proposed a novel method based on Hierarchical Convolutional Neural Networks for depression detection during the pandemic. The proposed method considered the user's history, and the attention mechanism learned the most important words and posts from the social media timeline. The model was trained on the dataset from Shen et al. [76] and achieved an 86.9% F1-score. The model was further used to identify users with depression using data from September 2019 to April 2020. The number of users predicted as depressed peaked in March when COVID-19 was declared a pandemic. LSTM and CNN-based neural network architectures were also used for depression detection during the COVID-19 pandemic [263–265].

Harrigian and Dredze [266] studied the semantic shift problem in longitudinal social media data monitoring during the COVID-19 pandemic. The authors used various depression datasets sampled from different social media platforms and time spans to measure semantic shift's influence on models for depression prevalence estimation. The classical approach for monitoring social media for depression is to train a statistical classifier on existing datasets and use the model to identify language associated with depression on new data collected from online platforms. Semantic shift, which occurs in longitudinal posts, can affect the performance of the models due to changes in the terminology for representing

concepts or in the language distribution [266]. The authors showed that semantic shift increased with the time gap between training and evaluation data collection. When training on historical data, small changes in the vocabulary can lead to large deviations for downstream monitoring tasks, even if these changes do not affect the performance of the models on the historical data. Furthermore, the authors used semantic shift measurements to curate vocabularies with semantic stability, improving statistical classifiers' predictive generalization. When comparing the most common social media platforms for COVID-19 monitoring, using vocabularies with semantic stability was more useful in the case of Twitter than on Reddit, due to the particularities of each social media platform. On Twitter, conversations focus on current events, while Reddit fosters discussion on specific interests (in subreddits) evolving over extended periods.

## 2.5 Datasets used in this Thesis

In this section, we present the mental-health-related datasets used in the thesis. We provide information on the annotation procedure, the platform where the data was collected, the number of users and posts in each dataset, and other relevant details. Our analysis focused on data from individuals who have received a formal diagnosis of mental disorders. Since obtaining this information can be challenging, we primarily depended on available datasets in which users explicitly mentioned mental health diagnoses in social media data, using these mentions as a proxy for formal diagnoses.

The **eRisk 2018 Depression** dataset [181] was created for the task of early detection of depression from social media timelines. The dataset is comprised of two groups: the depression and control groups. Individuals from the depression group were annotated by self-mentions of depression diagnoses. Only explicit mentions of a diagnosis were considered, such as *"I was diagnosed with depression*, while expressions such as *"I am depressed" or "I have depression"* were not considered. The mentions of diagnoses were removed from the dataset to prevent participants from easily identifying the users from the depression groups by these mentions. Users from the control group were collected from random individuals who did not mention the diagnosis in their posts. Control users from the depression subreddit[7] were also included in the control group. Thus, the dataset consists of user social media timelines and the corresponding binary label. The dataset was split into training and testing by the organizers of the eRisk Lab. The training dataset is made available to the participants so they can develop their methods. The testing data was released sequentially, so participants needed to make predictions on each post at a time before receiving more data. The training dataset contains 135 depressed users and 752 controls, while the test data has 79 depressed users and 741 controls. The dataset includes a total of 90,222 submissions from users with a

[7] https://www.reddit.com/r/depression/

depression diagnosis and 986,360 posts in the control group. The dataset was used in Chapters 3 and 4.

The **eRisk 2021 Pathological Gambling** dataset [88] was created for the early detection of pathological gambling from social media. It consists of users at risk of pathological gambling and control users, without any pathological gambling problems, from Reddit. Because 2021 was the first edition of the task, the organizers did not make available any training data. The test data was released sequentially, similar to the 2018 eRisk edition. The task consists of predicting if a given user is at risk of pathological gambling or not. The dataset contains 164 users at risk of pathological gambling, with a total of 54,674 posts. The control group includes 2,184 users with 1,073,883 posts. The dataset was used in Chapter 4.

The **eRisk 2021 Self-Harm** dataset [88] was created to detect users at risk of self-harm from Reddit. The dataset consists of users at risk of self-harm and control users who are not involved in these manifestations. The organizers provided both training and test data for this task. The training data is comprised of 145 users at risk of self-harm and 618 control users. The test data consists of 152 users from the self-harm group and 1,296 users from the control group. The dataset contains a total of 69,722 posts from individuals at risk of self-harm and 943,465 posts from the control group. The task consists of predicting the risk of self-harm for a given user. The dataset was used in Chapter 4.

The **eRisk 2022 Pathological Gambling** dataset [267] builds on the previous edition's data. The data from eRisk 2021 is considered the training dataset, while the test data was newly collected. The test data consisted of 81 users at risk of pathological gambling and 1,014,122 posts from the control group. The dataset was used in Chapter 4.

The **eRisk 2022 Depression** dataset [267] builds on the data collection from eRisk 2018, which was used as the training data. The test set contains 98 users with depression with 35,332 posts and 1,302 users in the control group with a total of 687,228 posts. The dataset was used in Chapters 4 and 6.

The **eRisk 2023** dataset for **Depression Symptoms Retrieval** [89] was created for ranking social media posts according to their relevance for the symptoms outlined in the BDI-II questionnaire [102]. The data was compiled from the eRisk past data and was organized as TREC formatted sentences for each user. A total of 3,807,115 sentences from 3,107 users were provided for this task. The dataset was used in Chapter 4

The **Reddit Self-reported Depression Diagnosis (RSDD)** dataset [74] consists of two groups of users, one with self-mentions of depression diagnosis and a control group, without any mention of a depression diagnosis. The dataset was collected from Reddit, and in the depression group, only users with explicit mentions of a diagnosis (e.g., *"I was diagnosed with depression"*) were included. To prevent users from the depression group from being easily identifiable, their mentions of diagnosis, mentions of depression-related terms (e.g., *depression, depressive*), or posts from mental health-related subreddits were removed. The dataset is split into

training, validation, and test sets, each with 3,070 users in the depression group and over 35,700 users in the control group. The dataset was used in Chapters 3, 4 and 6.

The dataset proposed by **Shen et al. [76]** is among the most widely used social media datasets in mental health research, as shown in Figure 2.3. The dataset was collected from Twitter, and users in the depression group were labeled by their mention of depression diagnosis. The dataset contains only social media posts posted one month after the diagnosis was mentioned. Individuals from the control group did not have any tweet containing the string *"depress"*, and their content was also collected in one month. We are using a version of the dataset similar to the one used in [164], which contains 1,402 users diagnosed with depression and 1,402 control users. It is a multimodal dataset, with the depression group containing a total of 213K tweets with only text and 19.3K tweets with text and images. The control group includes 828K text-only tweets and 50.6K tweets with text and images. This dataset was used in Chapter 5.

**MultiRedditDep** [64] is, to the best of our knowledge, the only multimodal dataset for depression detection sourced from Reddit. Similar to the other social media datasets, the depression group was annotated by mentions of depression diagnosis, while the control group was collected from users who did not have any mention of depression diagnosis. The control group is comprised of users active in the *r/depression* subreddit, but without any mention of depression diagnosis. The dataset is split into training, validation, and test sets, with 1,419 users in the depression group and 2,344 in the control group. The authors provided the train, validation, and test splits, with 2,633, 379, and 751 users, respectively. The dataset has 6.6M text-only posts and 46.9K posts with images in the depression group. The control group contains 8.1M text-only posts, and 73.4K posts with images. This dataset was used in Chapter 5.

**D-Vlog** [177] is a multimodal dataset from depression detection collected from YouTube vlogs. We consider this dataset an in-the-wild dataset, as it was collected from user-generated content with various recording conditions. It contains vlogs labeled as control or depression. The vlogs from the depression group were collected using depression-related keywords (e.g., "depression vlog"), while the videos from the control groups were collected using neutral keywords (e.g., "daily vlog"). The collected videos were manually reviewed, and vlogs from the depression group were considered only if they contained explicit mentions of depression symptoms. The dataset comprises 961 video samples with an average duration of 9.9 minutes, 555 vlogs in the depression group, and 406 videos in the control group. It is split into a training set with 647 samples, a validation set with 102 samples, and a test set with 212 samples. This dataset was used in Chapter 5.

**DAIC-WOZ** [268] is a dataset composed of clinical interviews recorded in controlled settings to diagnose mental disorders, such as anxiety, depression, and post-traumatic stress disorder. The interviews were conducted as Wizard-of-Oz interviews, where the interviewer was a virtual character named Ellie, who was controlled by a human interviewer. The dataset contains 189 sessions with an

average duration of 15.9 minutes. The data is highly imbalanced, containing 132 control and 57 subjects in the depression group. It is split into a training set with 107 samples, a validation set with 35 samples, and a test set with 47 samples. The authors of the DAIC-WOZ dataset provided the transcripts of the interviews alongside audio-based and video-based features. The ground truth is represented by the PHQ-8 [269] score, which is used to assess the severity of depression. This dataset was used in Chapter 5.

**E-DAIC** [270] is the extended version of DAIC-WOZ [268] and was collected following the same protocol. It contains a total of 275 sessions with an average duration of 16.2 minutes. Similar to DAIC-WOZ, it is severely imbalanced, containing 209 control and 66 subjects in the depression group. It is split into a training set with 163 samples, a validation set with 56 samples, and a test set with 56 samples. The E-DAIC dataset contains the transcripts of the interviews alongside audio-based and video-based features. Similar to DAIC-WOZ, we use the score from the PHQ-8 questionnaire is used as ground truth. This dataset was used in Chapter 5.

## 2.6 Summary

In this chapter, we presented how mental health is expressed through language, images, and online video content. We made an overview of the various methods used to detect these patterns in social media data, ranging from traditional machine learning models to recent advancements using large language models. Additionally, we provided a comprehensive review of natural language processing techniques developed to analyze depression in social media during the COVID-19 pandemic, while also offering insights from a post-pandemic perspective. Finally, we presented the mental health data collections used in this thesis.

CHAPTER 3

# Language Particularities of Depression in Social Media Data

In this chapter, we provide an extensive analysis of language particularities in the online discourse of social media users with depression based on our published papers:

- **A.-M. Bucur**, I. R. Podina and L. P. Dinu, *A Psychologically Informed Part-of-Speech Analysis of Depression in Social Media*, Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP) (2021) pp. 199–207
- **A.-M. Bucur**, M. Zampieri and L. P. Dinu, *An Exploratory Analysis of the Relation Between Offensive Language and Mental Health*, Findings of the Association for Computational Linguistics: ACL-IJCNLP (2021) pp. 3600–3606

Previous interdisciplinary research has highlighted some particularities of language use in individuals diagnosed with depression. For instance, individuals with depression may use part-of-speech tags differently than control individuals. In addition, previous research has shown that individuals with mood disorders tend to use more swear words in their online messages. In line with this research, we aim to conduct a comprehensive analysis of how individuals with a depression diagnosis use social media language.

In this chapter, we answer the following research questions:

**RQ1** Are there significant differences in part-of-speech usage in social media texts between individuals with a self-reported depression diagnosis and a control group?

**RQ2** Are there significant differences in the use of offensive language between individuals with self-reported depression diagnoses and control?

## 3.1 A Psychologically Informed Part-of-Speech Analysis of Depression in Social Media

Research in psychology revealed that individuals with depression tend to be self-focused, more preoccupied with themselves, and ruminate more about their lives and emotions. In this section, we aim to use large-scale datasets and computational methods to explore social media discourse quantitatively. We perform a part-of-speech (POS) analysis and contribute to the understanding of the differences in social media discourse between individuals from depression and control groups. We focus on the differences in part-of-speech use and ground them in the literature from psychology researchers. We aim to provide insights into the linguistic patterns of individuals with depression and see if patterns from psychology research are reflected in the social media discourse, which is more noisy and contains non-standard text [5]. We use the eRisk 2018 depression dataset and extract several part-of-speech features. Through this work, we aim to provide insights regarding how individuals with depression express themselves on social media, allowing for better-informed computational models to help monitor mental health problems.

### 3.1.1 Part-of-Speech Analysis

In our experiments, we use the dataset from the eRisk 2018 Lab [181] containing posts from individuals with self-mentions of a depression diagnosis and control users written in English from the social media platform Reddit. The dataset was presented in Section 2.5.

For each post in our dataset, we use the spaCy[1] part-of-speech tagger to extract the corresponding tags and morphological features (e.g., person and number for pronouns) for each word. We extract the universal POS tags[2] and the ones from The Penn Treebank tagset [271]. We use the latter to explore the differences in the verb tenses. We assign the tenses according to their tags: VBD and VBN corresponding to past tense, VBG, VBZ, and VBP corresponding to present tense, and MD tag before a VB corresponds to the future tense [272]. From the morphological features provided by the spaCy tagger, we extract the person and the number for all the pronouns. After this analysis, we use the following features in our exploration:

**Universal Part-of-Speech:** ADJ, ADV, NOUN, PROPN, VERB, ADP, CCONJ, DET, PART, SCONJ, AUX, PRON

**Verb tenses:** Past, present, future

**Person of pronouns:** First, second and third-person

**Pronoun number:** Singular and plural, only for the first-person

For each of these features, we compute their frequency for each post in the dataset. For the universal POS, the frequency is calculated as the number of occurrences of a specific tag normalized by the total number of tags in a post. For verb tenses, the

[1] https://spacy.io/

[2] https://universaldependencies.org/docs/u/pos/

frequency of each tense is calculated as a percentage of the total number of verb occurrences. For the three kinds of personal pronouns, each frequency is computed as a percentage of the number of all personal pronouns. The frequency of singular and plural first-person pronouns is calculated as a percentage of all first-person pronouns.

To further explore the part-of-speech usage in the social media dataset, we also use some other measurements based on POS tags [273]:

**Pronominalisation Index (PI):** reflecting the usage of pronouns, instead of another part-of-speech (e.g. nouns). It is computed as the number of pronouns divided by the number of nouns [274].

**Formality Metric** [275]:

$$F = \frac{NOUN + ADJ + PREP + ART - PRON - VERB - ADV - INTJ + 100}{2}$$

Moreover, we test the discriminatory power of POS tags frequency usage in users' texts and the specific computed indices. For this, we employ a Random Forest classifier on the train set of the eRisk 2018 dataset on the features mentioned above. To interpret the trained model and estimate each feature's importance, we employ SHapley Additive exPlanations (SHAP) [156] to measure each feature's contribution to the classifier decision.

**Classification** We opted for a simple Random Forest model, trained with 50 estimators and a max depth of 15, to avoid overfitting, with balanced class weights since the dataset is heavily imbalanced. On the test set for the eRisk 2018 dataset, we obtain a weighted $F_1$-score of 78.37% (with balanced class weights) and a macro $F_1$-score of 51.93%. While the classification task is difficult, we are interested in exploring the importance of the different features of the model, which sheds light on the model behavior and provides us with insights regarding which POS tag is most discriminatory.

### 3.1.2 Results and Discussion

In this section, we present our findings and results from our published work (Bucur et al. [3]). Addressing our research question, we compute the Welch t-test for all our features and demonstrate that there are statistically significant differences ($p$-value <0.001) in part-of-speech usage between the depression and control groups. This section presents these differences and their interpretation from the psychology research.

**Content Part-of-speech** In Figure 3.1, we present the usage of content words for the two groups from the eRisk 2018 dataset. Individuals diagnosed with depression tend to use fewer common and proper nouns in comparison with control users. They also use more verbs and adverbs in their posts. The discourse is focused around actions but with fewer entities (e.g., nouns), showing a defective linguistic structure with less interest in the environment (e.g., people, objects) [104].

To further understand these differences, we look at the frequencies of nouns and verbs in the social media discourse. We separately compute the keyness score [276,

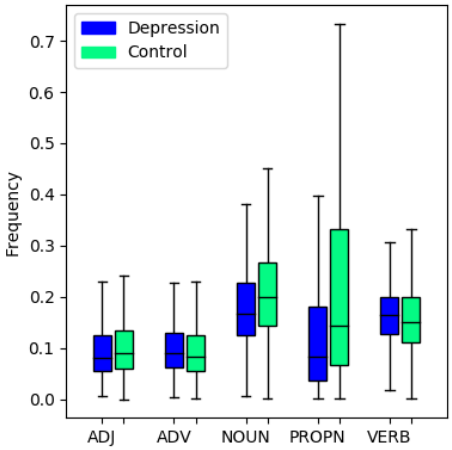



Figure 3.1: Frequency of content part-of-speech.

277] for verbs and nouns. In the keyness analysis, we compare the frequencies of nouns and verbs from the posts written by individuals in the depression group (target corpus) to the posts from the control group (reference corpus). In Figure 3.2, we present the top 20 verbs and nouns from each corpus, ordered by their log-likelihood ratio ($G^2$) [278].

Rumination is a cognitive process focusing on past and present negative content and resulting in emotional distress [63]. It is present in several mental health disorders (e.g., depression, anxiety, obsessive-compulsive disorder, post-traumatic stress disorder). In depression, rumination is defined as *"behaviors and thoughts that focus one's attention on one's depressive symptoms and the implications of these symptoms"* [279]. The rumination, as a response to depression, focuses the person's attention on their emotional state and inhibits the actions necessary to distract them from their mood. In Figure 3.3, when comparing the usage of the three verb tenses (present, past, and future) between the two groups, we expected that signs of rumination would be present in our analysis, with texts from users with depression being shifted into the past [47], but this result is not found in this sample of individuals.

Regarding future tense usage, individuals from the depression group have a lower frequency of verbs portraying future actions. This result may be a consequence of anhedonia, with people suffering from depression reporting lower anticipatory pleasure and thus talking online less about their future plans. Anhedonia, defined as *"markedly diminished interest or pleasure in all, or almost all activities most of the day, nearly every day"* [30], is a common symptom of depression.

The higher frequency of cognitive verbs (e.g., *feel, think, know*) in the texts written by individuals with depression indicates the cognitive impairments specific to depression. People with depression have cognitive deficits and biases in the processing of emotional information, and they are unable to regulate their emotions

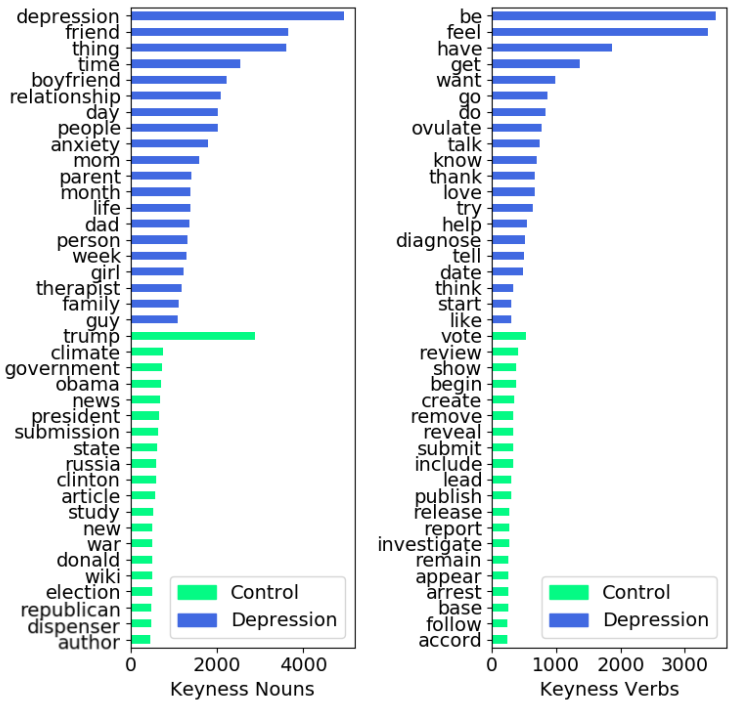



Figure 3.2: Most frequent nouns and verbs.

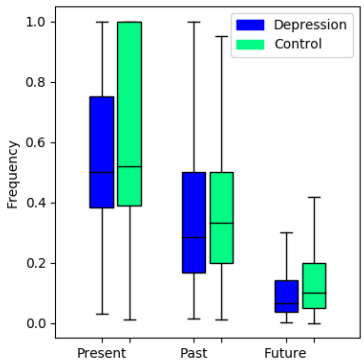



Figure 3.3: Frequency of verb tenses.

[280]. Individuals with or without depression may not differ in their initial response to an adverse event. Still, they differ in their ability to recover once they have experienced the negative emotion. In comparison, individuals from the control group use more action verbs (e.g., *vote, lead, show, begin, create*. Individuals with depression are more passive; they have a lower level of general activity, consistent with symptoms of depressive disorder [281].

Examining the frequencies of nouns in the eRisk 2018 dataset, we show in Figure

3.2 that the users with self-reported depression diagnoses use their Reddit account to disclose and discuss their mental health problems (e.g., *depression, anxiety, therapist*) or their personal relationships (e.g., *friend, boyfriend, relationship, mom, dad*). The process of seeking online support is also shown in the frequency of verbs: *feel, talk, diagnose, help*. This behavior of disclosing sensitive information on Reddit may stem from the pseudo-anonymity that the platform provides, allowing a high degree of information exchange related to mental health [280].

Even if the dataset contains control users active in the depression subreddit, most control users seem to post on other themed subreddits (e.g., politics). Bucur and Dinu [1] perform a topic modeling analysis and show that texts from control users are found in topics related to their hobbies, as opposed to individuals with depression, who are more focused on their feelings and life events. Our results align with this study; the users from the control group use more politics-related words (e.g., *trump, government, president, news, vote*).

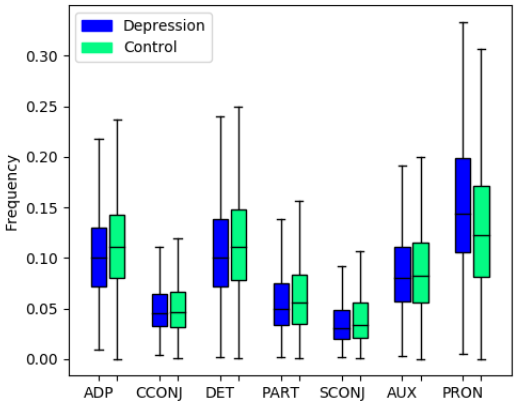



Figure 3.4: Frequency of functional part-of-speech.

**Functional part-of-speech** In Figure 3.4, we show the frequencies of functional parts of speech tags used by individuals in depression and control groups. Our findings indicate that individuals with depression tend to use fewer functional words in their writing than those in the control group, with the exception of pronouns. The reduced use of prepositions may be linked to decreased activation in the left hemisphere regions responsible for generating more abstract lexical units [274]. Function words serve an important social purpose and using them effectively requires strong social skills [282].

Figure 3.5 shows the differences in personal pronouns used by the two groups. The high frequency of first-person singular pronouns indicates a higher self-preoccupation in individuals with depression, as opposed to the first-person plural pronouns, which show collective attention. Second and third-person pronouns indicate social interactivity and contain references to other people or things in the environment [104].

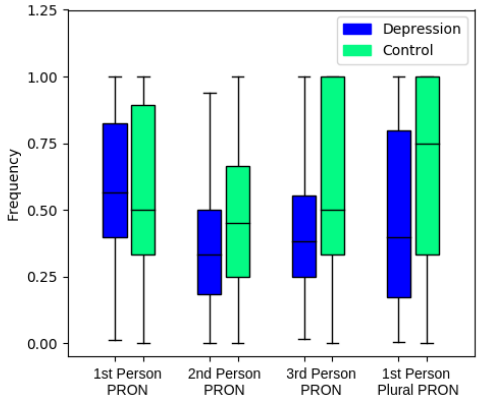



Figure 3.5: Frequency of personal pronouns.

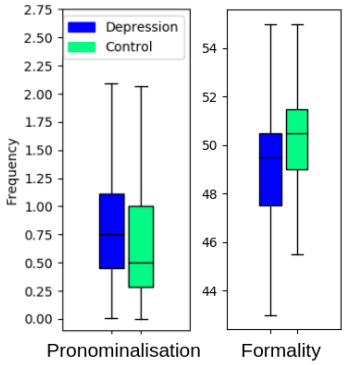



Figure 3.6: Indices based on part-of-speech tags.

Users in the depression group use more first-person singular pronouns, while the frequency of first-person plural pronouns is inversely related to the use of first-person singular pronouns. This result aligns with the Self-Focused Attention tendency (SFA) in individuals with depression. SFA is a cognitive bias commonly associated with depression; a high frequency of first-person singular pronouns in speech or written text is considered a linguistic marker of SFA [283]. Moreover, individuals with depression frequently struggle with social cognition that focuses on others. They often have impairments in Theory of Mind and empathy [284]. Theory of Mind allows individuals to infer the behaviors and intentions of others. Erle, Barth and Topolinski [285] show that individuals exhibiting high levels of depressive symptoms perform poorly on tasks involving overcoming their egocentrism.

The reduced use of first-person plural pronouns in texts written by individuals diagnosed with depression may reflect a decreased sense of belonging. People with depression often struggle with information-processing biases that hinder their ability to perceive cues of acceptance and belonging in social interactions. In addition, they are more likely to interpret ambiguous social situations negatively [286].

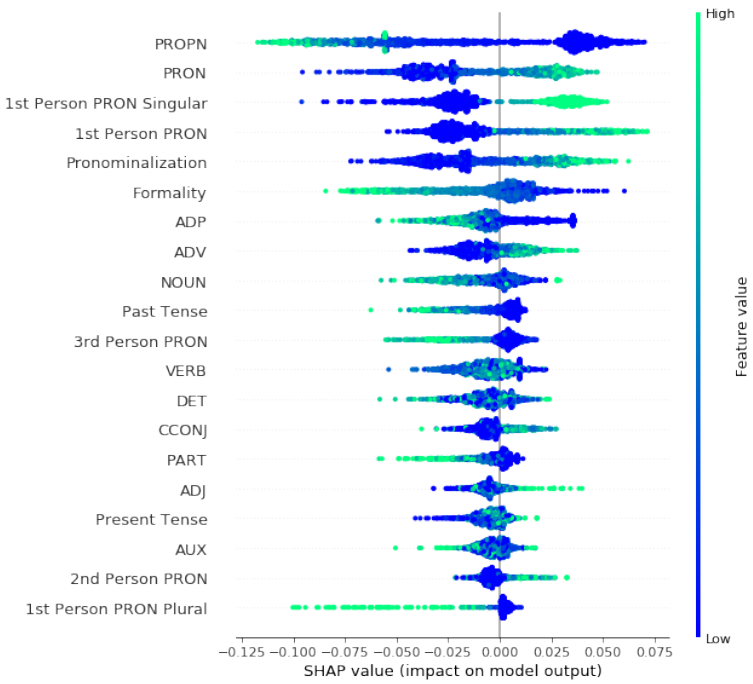



Figure 3.7: Feature importances of the RandomForest model for depression classification using POS frequencies and POS indices. As shown in our analysis, the usage of pronouns (1st Person Singular/Plural) and proper nouns is the most discriminatory.

**Pronominalisation and Formality Indices** In Figure 3.6, we present the results for the two metrics computed on part-of-speech tags. Individuals in the depression group have a higher pronominalisation index, using more pronouns despite nouns. In addition, they also use less formal language. They have a more contextualized discourse on social media, providing information about the context to avoid ambiguity [287].

Figure 3.7 shows Shapley values for a random sample of 1500 posts from the eRisk 2018 test set. A higher Shapley value corresponds with a higher importance in the final model decision based on the feature value. We used all computed POS features and indices in our model, but show only the top 20 for clarity. The absence of proper nouns is the most discriminatory factor in the decision to classify a person as part of the depression group. Moreover, using pronouns (also evident in the Pronominalisation Index) is highly correlated with positive model output. The high prevalence of first-person singular pronouns and the low prevalence of first-person plural pronouns support our findings from the exploratory analysis and psychology literature.

## 3.2 An Exploratory Analysis of the Relation Between Offensive Language and Mental Health

The prevalence of offensive language on social media has been widely studied. Offensive language can take various forms, from general and often harmless profanity to abusive language intended to cause harm, such as cyberbullying and hate speech [288]. Computational models have been applied not only to identify the various types of offensive content [289] but also to, for example, study the relation between profanity and hate speech [290] and the different functions and intentions of vulgarity in social media [291]. Most data used in the studies mentioned above is sampled from the general population, and there is a lack of research focused on studying offensive language use by specific groups, such as individuals with mental disorders. A notable exception is the recent study by Birnbaum et al. [48], which shows that users with mood disorders (bipolar disorder, and major depressive disorder) and schizophrenia spectrum disorders use more swear words in their Facebook messages than healthy users. To address this gap, we use available datasets created for offensive language identification and depression detection, and we train computational models to compare the use of offensive language in social media posts written by groups of individuals with and without self-reported depression diagnoses. We also look at samples written by groups of individuals whose posts show signs of depression. Through this contribution, we aim to open new avenues in research on offensiveness and mental health.

### 3.2.1 Offensive Language Detection and Categorization

To address **RQ2**, we first train a transformer-based model on a popular dataset for offensive language detection, and then we use it to label the posts from two social media datasets with posts from individuals with depression. In our experiments, we use three available English datasets with data collected from social media: one with offensive language labels and two datasets with posts from users with self-reported depression diagnoses.

**Offensive Language** We use the Offensive Language Identification Dataset (OLID) [292] to train offensive language identification models. OLID contains 14,100 manually annotated posts from Twitter, and it was released as the official dataset of SemEval-2019 Task 6 (OffensEval) [293]. We chose OLID due to its general hierarchical annotation taxonomy with the following levels:

**Level A:** Offensive language identification: offensive (OFF) vs. non-offensive (NOT)

**Level B:** Categorization of offensive language: targeted insult or threats (TIN) vs. untargeted profanity (UNT).

**Level C:** Offensive language target identification: individual (IND) vs. group (GRP) vs. other (OTH).

This hierarchical taxonomy provides us with flexibility as it represents multiple

types of offensive content in a single annotation scheme (e.g., posts targeted at an individual are often *cyberbullying*, and posts targeted at a group are usually *hate speech*) making it a great fit for this kind of analysis. In our experiments, we consider level A (offensive vs. non-offensive) and level B (target vs. untargeted).

**Mental Health** We run all our experiments on the RSDD dataset [74] and on the eRisk 2018 dataset [181]. The datasets were described in Section 2.5.

**Offensive Language Detection and Categorization** We address **RQ2** by studying the language of users from the two groups, self-reported depression diagnosis and control, in social media. We start by computing an offensive score, which measures the extent to which a post is offensive and whether it is a targeted insult or an untargeted post (most often profanity). These two tasks correspond to OLID levels A and B, respectively [292].

We fine-tune a BERT model on the OLID dataset on level A for offensive language detection. We train the model for two epochs, with a learning rate of 0.00001 and Adam optimizer [294]. We use an 80:20 split of the training data to choose the best-performing model in terms of F1 score. The model obtained 0.85 Precision, 0.74 Recall, and 0.77 F1 score on the test data from the OLID dataset. These numbers are consistent with the baselines reported in [292]. The offensive score is computed as a probability taken from the softmax output of the BERT model.

We also chose a transformer-based approach for offensive language categorization (targeted insult or untargeted profanity) using another BERT model trained on OLID level B. We fine-tune BERT for seven epochs with the same aforementioned train-validation split, with a learning rate of 0.00002 with Adam optimizer and a linear warm-up schedule with a 0.05 warm-up ratio, as proposed by Rosenthal et al. [295]. We use cross-entropy loss with balanced class weights to account for the class imbalance. The effectiveness of the model is also evaluated on the OLID test data, using the same metrics and achieving 0.78 Precision, 0.84 Recall, and 0.80 F1 score.

**Signs of Depression Detection** Furthermore, we are interested in distinguishing the posts that show signs of depression from all the posts of individuals from the depression group. This way, we filter out the noise added by the texts that do not contain any depression cues. We use the Semantic Polarity Score heuristic ($H_s$ heuristic) proposed by Ríssola, Bahrainian and Crestani [198] to detect posts showing signs of depression written by individuals with a self-reported depression diagnosis.

$H_s$ uses a mix of sentiment polarity, depression score, and emotion detection. The authors use TextBlob[3] to obtain the polarity score of each post, ranging between -1 and 1. The terms from EmoLex [296] are used to detect the eight Plutchik's emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) contained in the texts. The depression score of each post is computed using the NRC Affect Intensity Lexicon [297], ranging from 0 to 1. To distinguish the posts showing signs of depression from other posts of users with a self-reported depression diagnosis,

[3] https://textblob.readthedocs.io/en/dev/index.html

we follow the criteria from Ríssola, Bahrainian and Crestani [198]. Posts are labeled as showing signs of depression if the texts have a negative polarity, if sadness or disgust emotions are present, and if they have a depression score higher than 0.1.

### 3.2.2 Results and Discussion

In this section, we present the results of our work, as detailed in Bucur et al. [4].

Using the $H_s$ heuristic, we demonstrate that there is a statistically significant difference (Welch t-test, $p$-value <0.001) in terms of offensive language use between posts with signs of depression and posts that do not show any signs from individuals with self-reported depression diagnosis. Posts containing signs of depression have a higher offensive score than posts without any signs, in both eRisk 2018 and RSDD datasets, as shown in Figure 3.8.

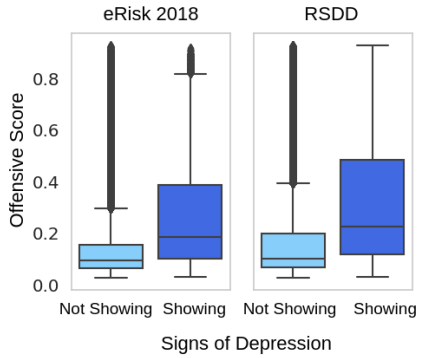



Figure 3.8: Distribution of the offensive language score for posts written by users with self-reported depression diagnosis and showing or not showing signs of depression measured with the $H_s$ heuristic.

We use the same 0.50 threshold used during training to label the offensive posts. We show in Table 3.1 that more posts from users diagnosed with depression are labeled as offensive than from control users. Using the $H_s$ heuristic, we filter the posts containing signs of depression and find that there is a higher percentage of posts with signs of depression labeled as offensive. These findings are consistent for both eRisk 2018 and RSDD datasets.

| | **Self-reported** | | **Signs of depression** | |
|---|---|---|---|---|
| **Dataset** | **Depression** | **Control** | **Showing** | **Not showing** |
| **eRisk 2018** | 8.24% | 5.91% | 18.50% | 7.40% |
| **RSDD** | 11.31% | 8.91% | 24.33% | 10.10% |

Table 3.1: Percentage of posts labeled as offensive from total posts of self-reported individuals and of posts showing/not-showing signs of depression measured with the $H_s$ heuristic.

The higher degree to which individuals with depression use offensive language in comparison to individuals in the control group can be explained via the emotion

regulation framework [298]. The use of offensive language could be an emotion regulation strategy through which individuals with depression relieve some of their distress. Similarly, studies on pain and distress indicate that the use of offensive language when experiencing pain significantly diminishes the level of pain experienced [299].

Although there are more posts with signs of depression labeled as offensive, the majority of them are untargeted (containing swears, profanity), and only 11.48% and 8.29%, respectively, are targeted insults (Table 3.2).

| | **Self-reported** | | **Signs of depression** | |
|---|---|---|---|---|
| **Dataset** | **Depression** | **Control** | **Showing** | **Not showing** |
| **eRisk 2018** | 24.12% | 21.72% | 11.48% | 26.68% |
| **RSDD** | 16.63% | 23.94% | 8.29% | 18.48% |

Table 3.2: Percentage of posts labeled as a targeted insult from the offensive posts of individuals with self-reported depression diagnoses and posts showing/not showing signs of depression measured with the $H_s$ heuristic.

The fact that individuals with depression tend to use more self-deprecating content and less deprecation of others, as evidenced in our analysis, is a result that is in line with the broad spectrum of cognitive studies, which indicates that negative evaluation of the self is a main interpretation bias in individuals with depression [300]. They tend to view themselves as less valuable than others. By self-deprecating language, we use the definition from [301]. This broader definition includes, but is not limited to, insults towards self, if they have a negative intention. Finally, studies show that there is also a self-focused attention tendency in depressed individuals [283], where just like in other conditions (e.g., anxiety), individuals tend to be unable to detach from their perspective, focusing primarily on their side of the story, their pain, etc.

To further understand the differences in the use of offensive language, we analyze the words from posts written by individuals with depression. We compute the keyness score [276, 277] of content words (removing stop words) from posts labeled as offensive written by users with self-reported diagnoses. The keyness is computed to show which words occur more often in the texts from individuals in the depression group showing signs of depression (target corpus) in comparison to the texts from users diagnosed with depression who do not show signs of depression (reference corpus). We calculate the frequencies of words from the two corpora and then the log-likelihood Ratio ($G^2$) [278] for each word. In Figure 3.9, we present the top 20 words, ordered by $G^2$ from each corpus, in the two datasets.

We show that, while posts without signs of depression refer more to sexual and profane terms, posts classified as showing signs of depression include more negative words such as *bad, hate, sick, death*. This result corroborates the findings described in the literature on cognitive errors or biases in depression [302]. It is well known that individuals with depression tend to view life events more negatively than the control peers [300]. Furthermore, individuals with depression are more

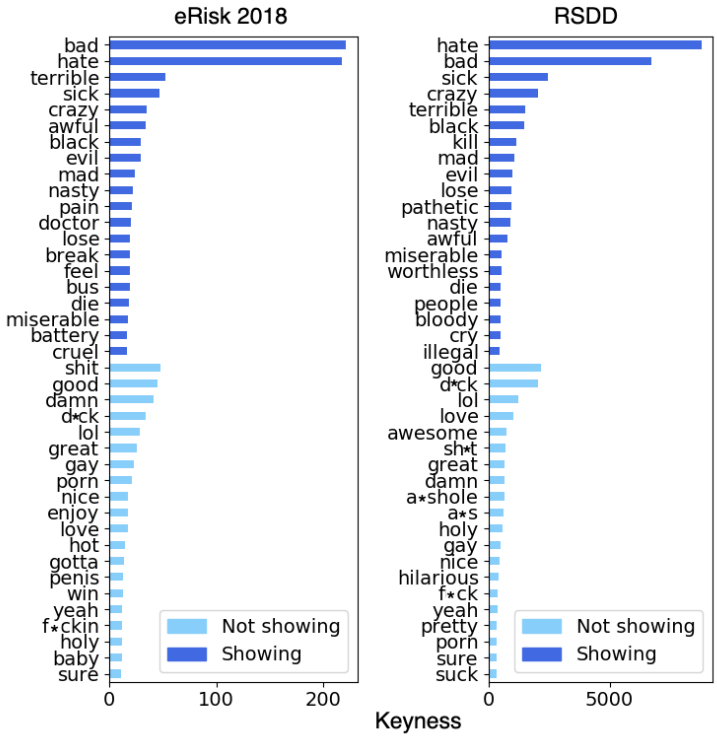



Figure 3.9: Keyness for words from posts showing/not showing signs of depression.

likely to recall negative life events than positive events and also more likely to pay closer attention to negative information [302]. Signs of this biased view of life are expected to be noticeable in language, and some studies indicate that individuals with depression tend to have a more negative discourse than the control peers [37]. Keywords with a negative polarity, such as *bad, die* or *pain*, seem to be pervasive in the speech of individuals with depression, as confirmed in our study. Finally, the reduced sexual drive is a well-known indication of depression [303]. Therefore, it is to be expected that individuals in the depression group tend to use fewer words with sexual connotations, as confirmed in our study.

## 3.3 Summary

In this chapter, we analyzed the language characteristics of social media users who mention having a depression diagnosis. To answer **RQ1**, we performed an extensive analysis of part-of-speech usage in social media texts from users with a depression diagnosis and control users [3]. Our findings were in line with studies in psychology and show that individuals diagnosed with depression use more pronouns (especially first-person singular pronouns) and verbs, as well as fewer common and proper nouns. Their social media discourse revolves around their life experiences and

sentiments, as opposed to control users who are not interested in discussing their problems online. Moreover, we also provided insights into the discriminatory power of POS frequencies by employing SHAP, a game-theoretic approach for model interpretation. We showed that users in the depression group can be characterized most easily, primarily through pronouns and proper nouns.

In addition, we applied offensive language identification techniques to posts by individuals with mental health conditions in order to analyze the use of profanity and offensive language by this group [4]. We showed how offensive language use differs substantially between individuals with depression, answering **RQ2**. Our findings indicate that users with self-reported depression diagnosis are more likely to use offensive language in their posts compared to the control group. From the posts of individuals with depression, the ones showing signs of depression contain more offensive language than the ones not showing any signs. In terms of the nature of offensive content, our results indicate that posts from individuals with signs of depression are less likely to contain targeted offensive language. Furthermore, while analyzing the texts of users with depression, we observed a larger frequency of words with negative polarity (e.g., *bad, hate, sick, suffer*) in the posts of users showing signs of depression, where the discourse of users not showing any signs contains more sexual-related content. These findings are consistent with the existing literature from psychology [299, 300, 302].

CHAPTER 4

# Early Risk Detection of Mental Disorders from Social Media Text

This chapter presents our contributions to developing early detection systems for mental disorders through our participation in the eRisk tasks from 2021 to 2023, and it is based on the following published papers:

- **A.-M. Bucur**, A. Cosma and L. P. Dinu, *Early Risk Detection of Pathological Gambling, Self-Harm and Depression Using BERT*, Working Notes of CLEF 2021, Early Risk Prediction on the Internet Lab (eRisk), CEUR 2936 (2021) pp. 938–949
- **A.-M. Bucur**, A. Cosma, L. P. Dinu and P. Rosso, *An End-to-End Set Transformer for User-Level Classification of Depression and Gambling Disorder*, Working Notes of CLEF 2022, Early Risk Prediction on the Internet Lab (eRisk), CEUR 3180 (2022) pp. 851–863
- **A.-M. Bucur**, *Utilizing ChatGPT Generated Data to Retrieve Depression Symptoms From Social Media*, Working Notes of CLEF 2023, Early Risk Prediction on the Internet Lab (eRisk), CEUR 3497 (2023) pp. 662–671
- **A.-M. Bucur**, *Leveraging LLM-Generated Data for Detecting Depression Symptoms on Social Media*, Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2024. Lecture Notes in Computer Science 14958 (2024) pp. 193–204

In this chapter, we aim to address the following research question:

**RQ3** How have recent advancements in artificial intelligence and natural language processing improved the early detection of mental disorders such as pathological gambling, self-harm, and depression, from social media texts?

## 4.1 eRisk 2021: Early Risk Detection of Pathological Gambling and Self-Harm Using BERT

In this section, we present our participation in the 2021 edition of the eRisk Lab, in which we tackle the problems of early detection of gambling addiction and self-harm (Bucur et al. [2]). According to the DSM-V, pathological gambling is a mental health illness in which a person has persistent and repetitive problematic gambling behavior. They are constantly preoccupied with gambling, feel a need to spend more money, and want to control or quit their gambling problem, but they already have several unsuccessful attempts; they try to hide their addiction and rely on others to regain a stable financial status or suffer other consequences [30]. Besides the classical gambling methods, with the rise of the Internet, the availability of gambling opportunities has increased. The incidence and prevalence of pathological gambling have also increased [35, 36].

Automated methods for compulsive gambling detection focus on analyzing behavioral player data (frequency of betting, number of different games played, etc.) [304, 305], transactional data [306] or texts from the correspondence between customers and customer service employees from online gambling operators [307]. To the best of our knowledge, the eRisk Lab is the first to tackle pathological gambling detection from social media text.

Self-harm is a serious mental health condition; people suffering from it are often diagnosed with other mental health disorders: borderline personality disorder, depression, anxiety, eating disorders, PTSD, and addictions (alcohol and drugs) [308]. Deliberate self-harm behavior refers to the act of self-inducing harm, resulting in tissue damage. It does not include other types of damage to the body (not exercising, unhealthy eating patterns, or engaging deliberately in actions resulting in psychological self-harm). Regarding adolescents, 17% had been involved in non-suicidal self-injury at least once. However, self-harm is less common in adults, with only 5% of them engaging in this behaviour.[1]

The task on early detection of signs of self-harm is a continuation of eRisk 2019's T2 and 2020's T1 tasks. The best-performing runs from the previous workshop edition in terms of Precision, $F1$, and Early Risk Detection Error ($ERDE$) were developed by Martínez-Castaño et al. [309]. The team used for sentence classification models based on XLM-RoBERTa [310] trained on different datasets. They avoided using the training dataset given by the eRisk organizers and instead gathered their own dataset from the Pushshift Reddit Dataset [311]. For the positive group, all posts from users who posted or commented in *r/selfharm* subreddit were extracted, including their submissions to other subreddits. The team manually annotated a small proportion of user posts, and the rest were used as positive samples without further labeling. For the control group, they extracted random users who had never posted on a self-harm-related subreddit. Other approaches to self-harm detection were using bag-of-words [312, 313], bag of sub-emotions [314], LIWC, emotion

[1] https://www.apa.org/monitor/2015/07-08/who-self-injures

and sentiment features [120] or deep learning models [183]. The results from this task suggest that the systems can obtain good performance on very little data and can be used to support human experts in the early detection of the signs of self-harm [183].

### 4.1.1 Early Detection of Signs of Pathological Gambling

The first task of this edition of the eRisk workshop is detecting early signs of pathological gambling. This is a novel task, and the organizers do not provide a training set; they only offer a testing platform. The data from this task is collected from Reddit, following the same methodology described by Losada and Crestani [108] and is described in Section 2.5. In 2019, the workshop switched from a chunk-based data release to a sequence-based approach for the test stage. Before 2019, the participants tested their runs on ten chunks with several submissions from each user, released weekly. In the current edition of the workshop, each team had to process their runs only on one post at a time from each user in chronological order.

Our approach is to construct a large dataset of user posts loosely annotated so that we can train our model. We follow a methodology similar to Martínez-Castaño et al. [309] and use the posts from the *r/GamblingAddiction*[2] and *r/problemgambling*[3] subreddits as positive training data for pathological gambling detection. The discussions in these subreddits primarily focus on gambling addiction. We consider the users who post in these subreddits to be at risk of pathological gambling. As such, we crawled the gambling-related subreddits and obtained 2,111 unique users with 13,377 posts only in these two subreddits. To avoid having only posts containing topics related to gambling, from the 2,111 users, we also crawled their posts from other unrelated subreddits, which increased the posts from the positive class to 80,998. For the collection of control users, we make use of the control split of the RSDD [74] and eRisk 2018 [108] depression datasets presented in Section 2.5.

Further, we fine-tuned a pre-trained BERT classifier using the AdamW optimizer with a learning rate of 0.0002 for two epochs. For prediction, we chose thresholds of 0.99 / 0.98 on the output probability to decide whether or not the user presents signs of gambling addiction. This high threshold minimizes false positives and ensures that a decision is taken only if a post denotes signs of addiction.

For this task, evaluation is performed through Precision, Recall, F1, the $ERDE$ [162] metric as an early risk detection measure, and $F_{latency}$ [131]. $F_{latency}$ was deemed more interpretable and informative for this task as it combines the effectiveness of the decision (estimated with the F measure) and the delay. Testing data for this task was released on an item-by-item basis through an API that processed current predictions for all runs and progressively released the following sequence data. We opted for five runs and processed 1,828 user writings for a total of 1 day and 23:43:28 hours.

[2] https://www.reddit.com/r/GamblingAddiction/
[3] https://www.reddit.com/r/problemgambling/

The five runs and the corresponding datasets used for classifying users at risk of pathological gambling are described below.

**RUN 0** BERT model trained on all the posts (including those from unrelated subreddits) from users who posted in the gambling-related subreddits, with control data from eRisk, with a 0.99 threshold.

**RUN 1** BERT model trained on all the posts (including those from unrelated subreddits) from users who posted in the gambling-related subreddits, with control data from RSDD, with a 0.99 threshold.

**RUN 2** BERT model trained on all the posts from the gambling subreddits with control data from eRisk, with a 0.99 threshold for classification.

**RUN 3** BERT model trained on all the posts (including those from unrelated subreddits) from users who posted in the gambling-related subreddits, with control data from eRisk, with a 0.98 threshold.

**RUN 4** BERT model trained on all the posts from the gambling subreddits with control data from eRisk, with a 0.98 threshold for classification.

Table 4.1: Decision-based evaluation on Task 1 of eRisk 2021: Early Detection of Signs of Pathological Gambling. **BLUE** represents the name of our team.

| team name | run id | P | R | F1 | $ERDE_5$ | $ERDE_{50}$ | $latency_{TP}$ | speed | latency-weighted F1 |
|---|---|---|---|---|---|---|---|---|---|
| **BLUE** | 0 | .107 | .994 | .193 | .067 | .046 | 2 | .996 | .192 |
| **BLUE** | 1 | .157 | .988 | .271 | .054 | .036 | 2 | .996 | .270 |
| **BLUE** | 2 | .121 | .994 | .215 | .065 | .045 | 2 | .996 | .215 |
| **BLUE** | 3 | .095 | **1** | .174 | .071 | .051 | 2 | .996 | .173 |
| **BLUE** | 4 | .110 | .994 | .198 | .068 | .048 | 2 | .996 | .197 |
| UNSL | 2 | **.586** | .939 | **.721** | .073 | **.020** | 11 | .961 | **.693** |
| RELAI | 0 | .138 | .988 | .243 | **.048** | .036 | 1 | **1** | .243 |

Table 4.2: Ranking-based evaluation on Task 1 of eRisk 2021: Early Detection of Signs of Pathological Gambling.

| | | 1 writing | | | 100 writings | | | 500 writings | | | 1000 writings | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| team | run | P@10 | NDCG@10 | NDCG@100 | P@10 | NDCG@10 | NDCG@100 | P@10 | NDCG@10 | NDCG@100 | P@10 | NDCG@10 | NDCG@100 |
| **BLUE** | 0 | .9 | .88 | .61 | .8 | .73 | .57 | .9 | .93 | .64 | .7 | .78 | .60 |
| **BLUE** | 1 | **1** | **1** | .61 | .8 | .82 | .53 | **1** | **1** | .56 | **1** | **1** | .56 |
| **BLUE** | 2 | .6 | .70 | .73 | .8 | .87 | .76 | .8 | .88 | .75 | .9 | .90 | .76 |
| **BLUE** | 3 | .6 | .65 | .60 | .8 | .87 | .61 | .7 | .71 | .60 | .7 | .67 | .60 |
| **BLUE** | 4 | .9 | .81 | .73 | **1** | **1** | .77 | **1** | **1** | .76 | **1** | **1** | .78 |
| UNSL | 2 | **1** | **1** | **.85** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** |

Our results, compared with the best results from the early detection of pathological gambling task, are presented in Table 4.1 for the decision-based evaluation and Table 4.2 for the ranking-based evaluation. Our runs obtained high Recall, but high Precision and F1 are also needed in a real-life scenario on problem gambling detection. Our $ERDE$ scores were relatively close to the best result from this task. Regarding the ranking-based evaluation, our team's ratings improved as our models analyzed more data from each user. Even after one writing from each user, the model was quite effective at predicting the risk of users and prioritizing them.

### 4.1.2 Early Detection of Signs of Self-Harm

For the second task, on early self-harm detection, the organizers provided a training dataset with users annotated at risk of self-harm and control users. The dataset was presented in Section 2.5. Given the small size of the data collection, we decided to increase the size of this dataset by crawling the *r/selfharm* subreddit. This subreddit contains various posts from users seeking help and sharing their experiences. We consider the users posting in this subreddit as being at risk of self-harm. We chose two variants for control users, one that uses the control split from RSDD and one that contains the control split from eRisk 2018, similar to the previous task on pathological gambling detection. We wanted to see the impact of the two different control datasets, as RSDD does not contain posts related to mental health, but eRisk does include posts from users active in mental health subreddits. The control splits used in our experiments contain random posts from users from various subreddits, and the number of posts was chosen to match the number of posts in the positive class.

We followed the same methodology as in the previous task. We also use a BERT model trained with the same parameters. For prediction, we chose thresholds of 0.99 / 0.98 / 0.95 on the output probability to decide whether or not the user presents signs of self-harm. This high threshold minimizes false positives and ensures that a decision is taken only if a post denotes signs of self-harm.

Similarly to the first task, testing data was released progressively after all runs offered predictions for the current batch of posts. Due to the lack of computational resources and server latency, we opted for five runs but processed only 156 user writings in 1 day and 4:57:23 hours.

Table 4.3: Decision-based results for Task 2 of eRisk 2021: Early Detection of Signs of Self-Harm.

| team name | run id | P | R | F1 | $ERDE_5$ | $ERDE_{50}$ | $latency_{TP}$ | speed | latency-weighted F1 |
|---|---|---|---|---|---|---|---|---|---|
| **BLUE** | 0 | .283 | .934 | .435 | .084 | .041 | 5 | .984 | .428 |
| **BLUE** | 1 | .142 | .875 | .245 | .117 | .081 | 4 | .988 | .242 |
| **BLUE** | 2 | .454 | .849 | .592 | .079 | .037 | 7 | .977 | .578 |
| **BLUE** | 3 | .394 | .868 | .542 | .075 | .035 | 5 | .984 | .534 |
| **BLUE** | 4 | .249 | .928 | .393 | .085 | .044 | 4 | .988 | .388 |
| Birmingham | 2 | **.757** | .349 | .477 | .085 | .07 | 4 | .988 | .472 |
| UPV-Symanto | 1 | .276 | .638 | .385 | **.059** | .056 | 1 | 1 | .0 .385 |
| UNSL | 0 | .336 | .914 | .491 | .125 | **.034** | 11 | .961 | .472 |
| UNSL | 4 | .532 | .763 | **.627** | .064 | .038 | 3 | .992 | **.622** |

The five runs and the corresponding datasets used to classify users at risk of self-harm are described below.

**RUN 0** BERT model trained only on the training data provided by the organizers, with a 0.95 threshold.

**RUN 1** BERT model trained on all the posts (including those from unrelated subreddits) from users who posted in the self-harm subreddit, with control data from eRisk, with a 0.99 threshold.

Table 4.4: Ranking-based evaluation for Task 2 of eRisk 2021: Early Detection of Signs of Self-Harm.

| | | 1 writing | | | 100 writings | | | 500 writings | | | 1000 writings | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| team | run | P@10 | NDCG@10 | NDCG@100 | P@10 | NDCG@10 | NDCG@100 | P@10 | NDCG@10 | NDCG@100 | P@10 | NDCG@10 | NDCG@100 |
| **BLUE** | 0 | 0.7 | 0.75 | 0.54 | 0.8 | 0.82 | 0.59 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **BLUE** | 1 | 0.2 | 0.13 | 0.26 | 0.4 | 0.41 | 0.29 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **BLUE** | 2 | 0.6 | 0.49 | 0.5 | **0.9** | **0.94** | 0.55 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **BLUE** | 3 | 0.6 | 0.43 | 0.49 | 0.8 | 0.87 | 0.54 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **BLUE** | 4 | 0.7 | 0.61 | 0.52 | 0.8 | 0.88 | 0.55 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| UPV-Symanto | 3 | 0.6 | 0.7 | 0.51 | **0.9** | **0.94** | 0.69 | 0.9 | 0.94 | 0.69 | 0.0 | 0.0 | 0.0 |
| UNSL | 0 | **1** | **1** | **0.7** | 0.7 | 0.74 | **0.82** | 0.8 | 0.81 | 0.8 | 0.8 | 0.81 | **0.8** |
| UNSL | 3 | **1** | **1** | 0.63 | **0.9** | 0.81 | 0.76 | 0.9 | 0.81 | **0.71** | 0.8 | 0.73 | 0.69 |

**RUN 2** BERT model trained on the training data provided by the organizers and posts from the self-harm subreddit, with control data from eRisk, with a 0.99 threshold.

**RUN 3** BERT model trained on the training data provided by the organizers and posts from the self-harm subreddit, with control data from eRisk, with a 0.98 threshold.

**RUN 4** BERT model trained on the training data provided by the organizers and posts from the self-harm subreddit, with control data from eRisk, with a 0.95 threshold.

Our results, along with the best results for the early detection task of self-harm, are presented in Table 4.3 for the decision-based evaluation and Table 4.4 for the ranking-based evaluation. Our team obtained good results in terms of F1 and $ERDE$ measures. For ranking-based assessment, we obtained the best P@10 and NDCG@10 scores after 100 writings among other teams, which means that these runs effectively detected users' self-harm risk.

## 4.2 eRisk 2022: An End-to-End Set Transformer for User-Level Classification of Depression and Gambling Disorders

This section presents our participation in the eRisk 2022 Lab. We propose a transformer architecture for user-level classification of gambling addiction and depression that is trainable end-to-end (Bucur et al. [7]). Unlike other methods that operate at the post level, we process a set of social media posts from a particular individual, to use the interactions between posts and eliminate label noise at the post level. We exploit the fact that, by not injecting positional encodings, multi-head attention is permutation invariant, and we process randomly sampled sets of texts from a user after being encoded with a modern pretrained sentence encoder (RoBERTa / MiniLM). Moreover, our architecture is interpretable with modern feature attribution methods and allows for automatic dataset creation by identifying discriminating posts in a user's text-set. We perform ablation studies on hyper-parameters and evaluate our method for the eRisk 2022 Lab on early detection of signs of pathological gambling and early risk detection of depression. The method proposed obtained the best $ERDE_5$ score of 0.015, and the second-best $ERDE_{50}$ score of 0.009 for pathological gambling detection. For the early detection of depression, we obtained the second-best $ERDE_{50}$ of 0.027.

For the task of detecting mental health disorders from social media text, many approaches focus on analyzing individual posts [2, 198, 315]. This assumes that if a user is experiencing depression, all of their posts will contain relevant information about this issue. However, we believe that this approach is inadequate, as many posts may be unrelated or uninformative regarding the user's mental health. Instead, the interactions among these posts could provide valuable insights into the user's mental well-being.

As such, we propose an architecture that performs user-level classification by processing a set of posts from a user. We exploit the fact that the multi-head attention operation in transformers is permutation invariant and inputs multiple texts from a single user into the model. This approach has several advantages: *(i)* it is trainable end-to-end, mitigating the need for hand-crafted construction of global user features *(ii)* it is robust to label noise, as some posts might be uninformative, the model learns to ignore them in the decision and *(iii)* it is interpretable, using feature attribution methods [316] we can extract the most important posts for the decision.

**Pathological Gambling** In the first iteration of the task for gambling addiction detection, the best-performing systems were developed by Maupomé et al. [317] and Loyola et al. [318]. Maupomé et al. [317] used a user-level approach based on the similarity distance between the vector of topic probabilities of the users' texts to be assessed for pathological gambling risk and testimonials or items from a self-evaluation questionnaire for gambling addiction. By using this method, the authors obtain the best $ERDE_5$ of 0.048. Loyola et al. [318] attain the best $ERDE_{50}$ (0.020) and latency-weighted F1 (0.693) through a post-level rule-based early alert

policy on bag-of-words text representation classified with Support Vector Machines (SVM).

**Depression** For the task of early detection of depression, the best systems from the first iteration of the task (eRisk 2017) used as input linguistic information extracted from the texts such as LIWC [121], readability and hand-crafted features [319] obtaining the best $ERDE_5$ (12.70%) or a combination of linguistic information and temporal variation of terms from users' posts [320] achieving the best $ERDE_{50}$ (9.68%). The best-performing systems from eRisk 2018 were the ones from Funez et al. [321] and Trotzek, Koitka and Friedrich [322]. Funez et al. [321] propose a user-level approach using an SVM classifier on semantic representations that consider the temporal variation of terms between the users' posts and achieve an $ERDE_5$ of 8.78%. On the other hand, the best $ERDE_{50}$ (6.44%) is attained by Trotzek, Koitka and Friedrich [322] using a chunk-level[4] approach using an ensemble of logistic regression classifiers on bag-of-words features. The dataset for the depression detection task from the eRisk Lab was an important resource later used in different research articles tackling the detection problem using approaches such as a neural network architecture on topic modeling features [1], SVM or deep learning architectures using fine-grained emotions features [323] or deep learning methods using content, writing style and emotion features [154].

### 4.2.1 End-to-End Transformer Architecture

As proposed by Vaswani et al. [324], the transformer encoder consists of multiple sequential layers of multi-head attention. Multi-head attention is *permutation invariant*, and the current way to inject temporal information into the input sequence is by employing positional encodings [325]. This is useful when processing sequential data such as texts. However, by omitting positional encodings, the transformer essentially acts as a *set encoder*. Lee et al. [326] introduced the Set Transformer, in which they prove that multi-head attention is permutation invariant and that the Set Transformer is a universal approximator of permutation invariant functions. We use this fact to perform user-level classification by processing *sets of texts* (in the form of social media posts) from a particular user. The intuition behind processing a set of texts from a user is that no single social media post is sufficiently informative for a classifier decision, but rather their interaction and the user behavior as a whole. Moreover, through mean pooling, the inevitable noise (in terms of unrelated posts) is dampened, which aids classification in weakly-supervised scenarios, such as ours, in which a user is labeled rather than all of their posts.

We consider a user $i$ to contain multiple social media posts $U_i$. A set of $K$ texts $t$ are randomly sampled from $U_i$, which defines our text-set $S_i = \{t^j \sim U_i, j \in (1 \dots K)\}$. We sample $K$ posts from the user's history, instead of processing all of them due

[4] In 2018, the test data was released in chunks of posts, not one post at a time as it is the case in this task

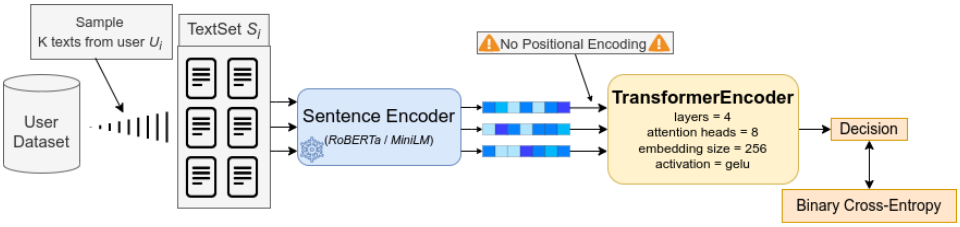



Figure 4.1: Proposed model architecture for eRisk 2022. We perform user-level classification by operating on a sample of K texts from a user. Texts are encoded with a pretrained sentence encoder and processed by a permutation-invariant transformer model. Binary cross-entropy loss is applied at the user level for a text-set.

to memory limitations - some individuals have thousands of posts while others have only in the order of tens. Moreover, stochasticity is introduced in the training procedure, which prevents overfitting. As such, for training, an input batch of size $n$ is defined by the concatenation of $n$ such text-sets: $B = \{S_{b_1}, S_{b_2}, \ldots S_{b_n}\}$. We do not consider the relative order of the texts for a particular user, and text-sets are fed into the transformer encoder without using positional encoding. Since some users have a total number of texts smaller than $K$, creating a batch of text-sets is only possible with padding and masking. However, to alleviate this problem, we train with an effective batch size of 1 and employ gradient accumulation to simulate a larger batch size.

Figure 4.1 showcases our proposed model architecture for user-level classification. Each text in a text-set is embedded into a fixed-size vector using available pretrained sentence encoder models (i.e., RoBERTa / MiniLM). The text embeddings are fed into the transformer encoder, and after processing, we perform mean pooling and output the decision. We compute binary cross-entropy at the user level for a text-set. The pretrained sentence encoder is frozen and not updated during training. We process social media posts directly as a set. The main reason for considering the posts as a set is that in a user's post history, many posts are uninformative to the modeling task, and by processing a set of texts, label noise is reduced naturally as a direct consequence of the attention mechanism, which assigns more importance to informative posts. However, training with a sufficiently large dataset might achieve the same effect, but previous attempts at post-level classification have proven ineffective [2].

To assess the impact of the sentence representations, we chose two different sentence encoders: RoBERTa [327] and MiniLM [328]. We chose RoBERTa since it is one of the best performing English language models in downstream tasks [327], and MiniLM, a multilingual model, since some users have social media posts in languages other than English. Figure 4.2 showcases the performance gap between the two sentence encoders, averaged across multiple values of $K$. RoBERTa yields a consistently better performance across training steps. Similarly, to assess the impact of the text-set size $K$, we performed an ablation study, as shown in Figure 4.3. We

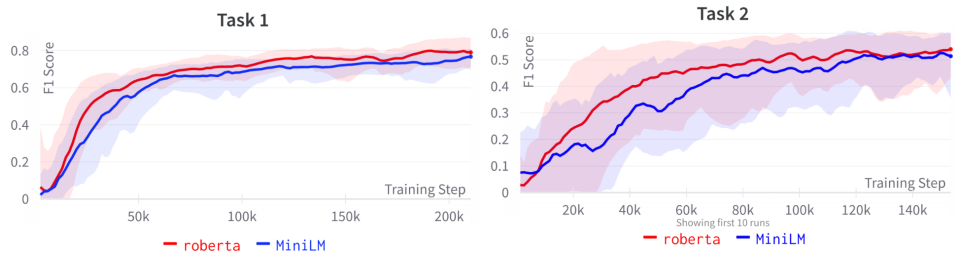



Figure 4.2: Performance of our model across training steps, in terms of $F_1$ score, for different sentence encoders (RoBERTa / MiniLM). We show the mean and standard deviation of $F_1$ score across multiple values of $K$. For both tasks, RoBERTa yields consistently superior performance compared to MiniLM.




Figure 4.3: Performance of our model across training steps, in terms of validation $F_1$ score, for RoBERTa sentence embeddings and varying the $K$, the number of texts per user. Tasks 1 and 2 have the best performance with $K = 16$ and $K = 32$, respectively.

kept the sentence encoder fixed to RoBERTa, and varied the number of texts per user $K \in \{4, 8, 16, 32, 64, 128\}$. The best performance was achieved with $K = 16$ and $K = 32$ for pathological gambling and depression detection, respectively.

In our final submission, we chose RoBERTa as a sentence encoder and sampled $K = 16$ texts per user for Task 1 and $K = 32$ for Task 2. We used the standard formulation of the transformer model [324]. Both models were trained for 120 epochs, with AdamW optimizer [294], with a cyclical learning rate [329] ranging from 0.00001 to 0.0001 across six epochs and a batch size of 128. To account for class imbalance, we computed balanced class weights with respect to each dataset and adjusted the loss function accordingly. Finally, we opted for a very high threshold when predicting the final decision.

Our proposed architecture can be easily interpretable using modern explainability methods for feature attribution [156, 157, 316], such as Integrated Gradients [316]. It automatically identifies social media posts containing signs of mental health disorders and filters out uninformative posts.

### 4.2.2 Interpretability

Since our model operates on sets of social media texts from a particular user, we can employ model explainability methods to assess the importance of a piece of text to the model decision. Through this, automatic filtering and selection of the most indicative posts of a user can be made for use in dataset creation. This idea is similar to Ríssola et al. [198], which employed a series of heuristics to recognize posts portraying depression symptoms to construct a post-level training set from existing depression datasets annotated at the user level. As such, we use Integrated Gradients [316] to compute attribution scores for a text-set. The integrated gradients method has been used in NLP to explore the contribution of individual words and phrases to a decision made by a classifier. Since we are not operating on words but instead on whole texts, this method computes the most important texts for the classifier decision.

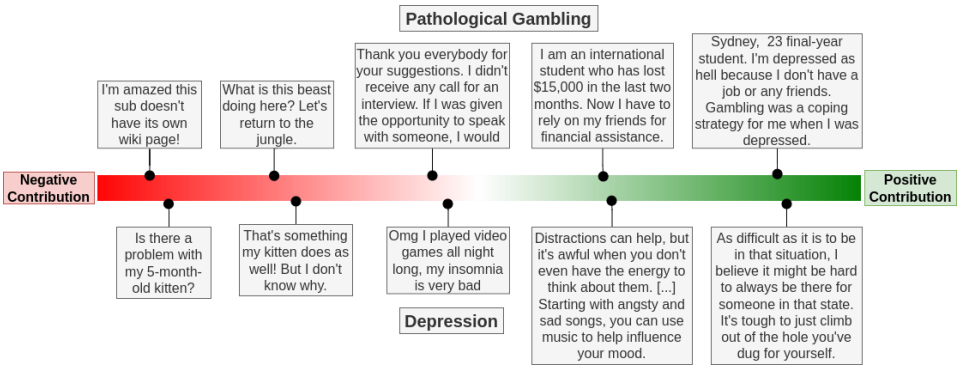



Figure 4.4: Texts from a particular user, relatively ranked in terms of attribution scores (contribution to a positive decision by the model) computed with the Integrated Gradients method. For each task, all texts belong to a single text-set of a user. The model can identify posts with clear discriminating information for each task. Examples have been paraphrased for anonymity.

Figure 4.4 showcases selected samples ordered by their attribution score from the validation set of each task. All samples belong to the same user for each task, and the attribution scores are computed within the respective text-set. Posts with a high positive contribution to the decision contain more explicit descriptions of symptoms, while posts with more negative contributions are mainly unrelated to the particular mental illness. We use the integrated gradients method in one of our runs to select the most important posts in the user history.

### 4.2.3 Results

#### Evaluation

There are two kinds of evaluation used for measuring the performance of the systems: decision-based and raking-based. The **decision-based evaluation** is

used for quantifying the capacity of a system to perform the binary classification and predicting if a user is from the positive class (i.e., pathological gambling or depression) or the negative one. It is comprised of standard measures for classification (Precision, Recall, F1) and measures for this specific task of early detection that consider the delay and the speed of the decision. $ERDE$ [108] measures the correct predictions considering a late decision penalty (for predictions taken after the 5 or 50 first submissions of a user). To overcome the limitations of this metric [330], the *latency-weighted F1* score [131] was also proposed to measure the performance of early risk detection. *Latency* measures the delay in detecting true positives based on the median number of submissions seen by the system before taking a decision. The *speed* of a system that correctly predicts true positives from the first submission is equal to 1, while a slow system decides after processing hundreds of texts. The latency-weighted F1 combines the F1 score with the delay in decision-making for true positives. A perfect system should achieve a latency-weighted F1 of 1. Besides the binary classification decisions, the participating teams were also asked to submit a score for estimating the risk of users for the **ranking-based evaluation**. These scores are used to rank users' risk for pathological gambling or depression. Standard IR metrics (P@10, NDCG@10, and NDCG@100) are used to measure the models' ranking-based performance after processing 1, 100, 500, or 1000 submissions.

### Task 1: Early Detection of Signs of Pathological Gambling

The first task proposes the detection of gambling addiction from social media data. This being the second edition of this task, the organizers provided last year's test data for training the systems, which is described in Section 2.5. For the testing phase, the submissions of users were released sequentially, and the systems proposed by the participating teams received one submission at a time from all the users. We submitted three runs for the early detection of pathological gambling: **Run 0** is comprised of the text-set transformer model using the most recent $K = 16$ posts for prediction; the system for **Run 1** is the same text-set transformer model using as input the set of $K = 16$ texts that are most important in a user's history, selected with Integrated Gradients; **Run 2** is a baseline run, using the proposed model architecture for predicting at post-level, on one sample at a time.

Table 4.5 showcases the performance of the systems measured using the decision-based measures. Regarding $ERDE$, our first run (**Run 0**), using the transformer architecture on the most recent texts from each user, manages to achieve the best $ERDE_5$ score of 0.015, and the second-best $ERDE_{50}$ score of 0.009, demonstrating that the system could detect early the true positive cases. The scores for $latency_{TP}$ and *speed* show that our models were successful at detecting the true positive cases after the first writing. As expected, the baseline run using a post-level approach (**Run 2**) has the lowest performance. Furthermore, our best run from eRisk 2022 surpasses all the runs from our participation in the first iteration of the task in 2021 [2], showing that a user-level approach considering a set of texts from each

Table 4.5: Decision-based evaluation on Task 1 of eRisk 2022: Early Detection of Signs of Pathological Gambling. We show the performance of our systems compared to the best-performing run from each team. **BLUE** represents the name of our team.

| Team | Run ID | P | R | F1 | $ERDE_5$ | $ERDE_{50}$ | $Latency_{TP}$ | Speed | Latency-Weighted F1 |
|---|---|---|---|---|---|---|---|---|---|
| **BLUE** | 0 | 0.260 | 0.975 | 0.410 | **0.015** | 0.009 | **1.0** | **1.000** | 0.410 |
| **BLUE** | 1 | 0.123 | 0.988 | 0.219 | 0.021 | 0.015 | **1.0** | **1.000** | 0.219 |
| **BLUE** | 2 | 0.052 | **1.000** | 0.099 | 0.037 | 0.028 | **1.0** | **1.000** | 0.099 |
| UNED-NLP | 4 | 0.809 | 0.938 | **0.869** | 0.020 | **0.008** | 3.0 | 0.992 | **0.862** |
| SINAI | 1 | 0.575 | 0.802 | 0.670 | **0.015** | 0.009 | **1.0** | **1.000** | 0.670 |
| BioInfo_UAVR | 4 | 0.192 | 0.988 | 0.321 | 0.033 | 0.011 | 5.0 | 0.984 | 0.316 |
| RELAI | 2 | 0.052 | 0.963 | 0.099 | 0.036 | 0.029 | **1.0** | **1.000** | 0.099 |
| BioNLP-UniBuc | 4 | 0.046 | **1.000** | 0.089 | 0.032 | 0.031 | **1.0** | **1.000** | 0.089 |
| UNSL | 1 | 0.461 | 0.938 | 0.618 | 0.041 | **0.008** | 11.0 | 0.961 | 0.594 |
| NLPGroup-IISERB | 4 | **1.000** | 0.074 | 0.138 | 0.038 | 0.037 | 41.5 | 0.843 | 0.116 |
| stezmo3 | 4 | 0.160 | 0.901 | 0.271 | 0.043 | 0.011 | 7.0 | 0.977 | 0.265 |

Table 4.6: Ranking-based evaluation on Task 1 of eRisk 2022: Early Detection of Signs of Pathological Gambling.

| | | 1 writing | | | 100 writings | | | 500 writings | | | 1000 writings | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Team | Run ID | P@10 | NDCG@10 | NDCG@100 | P@10 | NDCG@10 | NDCG@100 | P@10 | NDCG@10 | NDCG@100 | P@10 | NDCG@10 | NDCG@100 |
| **BLUE** | 0 | **1.00** | **1.00** | **0.76** | **1.00** | **1.00** | 0.81 | **1.00** | **1.00** | 0.89 | **1.00** | **1.00** | 0.89 |
| **BLUE** | 1 | **1.00** | **1.00** | **0.76** | **1.00** | **1.00** | 0.89 | **1.00** | **1.00** | 0.91 | **1.00** | **1.00** | 0.91 |
| **BLUE** | 2 | **1.00** | **1.00** | 0.69 | **1.00** | **1.00** | 0.40 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.01 |
| UNED-NLP | 4 | **1.00** | **1.00** | 0.56 | **1.00** | **1.00** | 0.88 | **1.00** | **1.00** | **0.95** | **1.00** | **1.00** | **0.95** |
| UNSL | 0 | **1.00** | **1.00** | 0.68 | **1.00** | **1.00** | **0.90** | **1.00** | **1.00** | 0.93 | **1.00** | **1.00** | **0.95** |

individual is more suitable than a post-level approach. In Table 4.6, we show the results of the ranking-based evaluation, in which each team had to submit the rankings of users' risk for pathological gambling. Our team achieves high results for NDCG and P@10 in all situations (after 1, 100, 1000, and 5000 writings).

### Task 2: Early Detection of Depression

This task marks the third iteration of the early detection of depression task, continuing the 2017 T1 and 2018 T2 tasks. The organizers provided the data from the previous two editions to train the models, which was described in Section 2.5. The texts for making the predictions for the testing phase were released sequentially, and the systems from the participating teams had to decide whether to make a decision for a specific user or wait for more data. We submitted three runs for the early detection of depression: **Run 0** is the text-set transformer model using the most recent $K = 32$ posts for prediction; for **Run 1**, we employ the same text-set transformer model using as input the set of $K = 32$ texts that are most important in a user's history, selected with Integrated Gradients; **Run 2** is a baseline run, using the proposed model architecture for predicting at post-level, on one sample at a time.

In Table 4.7, we present the performance of the systems using the decision-based

Table 4.7: Decision-based evaluation on Task 2 of eRisk 2022: Early Detection of Depression. We show the performance of our systems compared to the best-performing run from each team.

| Team | Run ID | P | R | F1 | $ERDE_5$ | $ERDE_{50}$ | $Latency_{TP}$ | Speed | Latency-Weighted F1 |
|---|---|---|---|---|---|---|---|---|---|
| **BLUE** | 0 | 0.395 | 0.898 | 0.548 | 0.047 | 0.027 | 5.0 | 0.984 | 0.540 |
| **BLUE** | 1 | 0.213 | 0.939 | 0.347 | 0.054 | 0.033 | 4.5 | 0.986 | 0.342 |
| **BLUE** | 2 | 0.106 | **1.000** | 0.192 | 0.074 | 0.048 | 4.0 | 0.988 | 0.190 |
| CYUT | 0 | 0.165 | 0.918 | 0.280 | 0.053 | 0.032 | 3.0 | 0.992 | 0.277 |
| LauSAn | 4 | 0.201 | 0.724 | 0.315 | **0.039** | 0.033 | **1.0** | **1.000** | 0.315 |
| BioInfo_UAVR | 4 | 0.378 | 0.857 | 0.525 | 0.069 | 0.031 | 16.0 | 0.942 | 0.494 |
| TUA1 | 4 | 0.159 | 0.959 | 0.272 | 0.052 | 0.036 | 3.0 | 0.992 | 0.270 |
| NLPGroup-IISERB | 0 | 0.682 | 0.745 | **0.712** | 0.055 | 0.032 | 9.0 | 0.969 | **0.690** |
| RELAI | 0 | 0.085 | 0.847 | 0.155 | 0.114 | 0.092 | 51.0 | 0.807 | 0.125 |
| UNED-MED | 1 | 0.139 | 0.980 | 0.244 | 0.079 | 0.046 | 13.0 | 0.953 | 0.233 |
| Sunday-Rocker2 | 1 | 0.355 | 0.786 | 0.489 | 0.068 | 0.041 | 27.0 | 0.899 | 0.439 |
| SCIR2 | 3 | 0.316 | 0.847 | 0.460 | 0.079 | **0.026** | 44.0 | 0.834 | 0.383 |
| UNSL | 2 | 0.400 | 0.755 | 0.523 | 0.045 | 0.026 | 3.0 | 0.992 | 0.519 |
| E8-IJS | 0 | **0.684** | 0.133 | 0.222 | 0.061 | 0.061 | 1.0 | 1.000 | 0.222 |
| NITK-NLP2 | 3 | 0.149 | 0.724 | 0.248 | 0.049 | 0.039 | 2.0 | 0.996 | 0.247 |

Table 4.8: Ranking-based evaluation on Task 2 of eRisk 2022: Early Detection of Depression.

| | | 1 writing | | | 100 writings | | | 500 writings | | | 1000 writings | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Team | Run ID | P@10 | NDCG@10 | NDCG@100 | P@10 | NDCG@10 | NDCG@100 | P@10 | NDCG@10 | NDCG@100 | P@10 | NDCG@10 | NDCG@100 |
| **BLUE** | 0 | **0.80** | **0.88** | **0.54** | 0.60 | 0.56 | 0.59 | 0.80 | 0.81 | 0.66 | **0.80** | 0.80 | 0.68 |
| **BLUE** | 1 | **0.80** | **0.88** | **0.54** | 0.70 | 0.64 | **0.67** | 0.80 | 0.84 | **0.74** | **0.80** | **0.86** | **0.72** |
| **BLUE** | 2 | **0.80** | 0.75 | 0.46 | 0.40 | 0.40 | 0.30 | 0.30 | 0.35 | 0.20 | 0.30 | 0.38 | 0.16 |
| NLPGroup-IISERB | 0 | 0.00 | 0.00 | 0.02 | **0.90** | 0.92 | 0.30 | **0.90** | **0.92** | 0.33 | 0.00 | 0.00 | 0.00 |
| Sunday-Rocker2 | 1 | 0.70 | 0.81 | 0.39 | **0.90** | **0.93** | 0.66 | **0.90** | 0.88 | 0.65 | 0.00 | 0.00 | 0.00 |

metrics. Our best-performing run is the transformer architecture using the most recent texts from users (**Run 0**), followed by the system that considers only the most informative submissions from each user for the model's decisions (**Run 1**). The post-level system (**Run 2**) has the worst performance. Our three submitted runs achieve high Recall at the expense of lower Precision scores. The precision of our models can be improved by incorporating a mechanism for weighting user posts according to the prevalence of signs of depression [198]. Regarding the early detection evaluation, our team has the second-best score on the $ERDE_{50}$ metric (0.027), while our $ERDE_5$ score is close to the best one. Compared to the best metrics from the 2018 edition of this task, when the best $ERDE_5$ and $ERDE_{50}$ were 0.087 and 0.064, respectively, current systems surpass these scores due to more data being available for training the models and the advancements in the field of machine learning. Regarding the standard metrics for classification, a slight improvement was made in terms of F1 score, from 0.64 in 2018 to 0.71 in 2022. The ranking-based evaluation performance from Table 4.8 shows that for 1 and 1000 writings, our systems attain some of the best scores for P@10 and NDCG.

## 4.3 eRisk 2023: Leveraging LLM-generated Data for Detecting Depression Symptoms on Social Media

In this section, we present our contributions to the eRisk 2023 task focused on identifying symptoms of depression in Reddit social media posts (Bucur et al. [10, 18]). The task consists of retrieving and ranking Reddit social media sentences that convey symptoms of depression from the BDI-II questionnaire. We used synthetic data generated by GPT-3.5 and Llama-3 for each BDI-II symptom to augment our data and improve downstream models. We used semantic search and cosine similarity to rank the relevance of the sentences to the BDI-II symptoms. We compared the performance of two transformer-based models (MentalRoBERTa and a variant of MPNet) in embedding social media posts and the original and generated BDI-II responses for information retrieval.

Our proposed method for searching symptoms of depression, was inspired by recent works on generating and augmenting data using LLMs [146, 149]. We follow a similar approach, and generate synthetic Reddit posts for each of the BDI-II symptoms such that the generated data has more diversity than the responses from BDI-II. We hypothesize that by generating synthetic data similar to the BDI-II responses with GPT-3.5, we will add more diversity to the data and will be able to retrieve more relevant sentences. We aim for the generated data to resemble Reddit posts, in which users share their experiences more intimately.

### 4.3.1 Retrieving Symptoms of Depression

The task from the eRisk 2023 Lab [89] consists of ranking sentences from social media posts according to their relevance to the symptoms from BDI-II [102]. The BDI-II is a questionnaire used by mental health professionals to screen for depression and consists of 21 questions related to symptoms of depression, such as sadness, pessimism, loss of pleasure, loss of interest, tiredness, and others. Each question corresponds to one of the symptoms. BDI-II is a Likert scale survey; for each question, there are four possible responses measuring the intensity of the symptom from the absence of it to its maximum intensity (with the exception of items 16 and 18[5], which have seven possible responses). The challenge consists of ranking the sentences from Reddit by their relevance to each of the symptoms in the BDI-II. A given sentence is considered relevant to a symptom if it contains information about the user's mental state regarding the symptom, even if the user mentions that they do not suffer from the given symptom. The data provided for the task was presented in Section 2.5.

For evaluating the systems' performance and assessing the sentences' relevance to the BDI-II symptoms, top-k pooling was used, with k equal to 50. The top 50 relevant sentences for each symptom from each system were combined in a pool of relevant sentences. These sentences were further assessed as being relevant or

[5] Related to changes in sleeping patterns and appetite.

not to the symptoms by three annotators. A sentence was considered relevant to a symptom if it contained information about the state of the individual and was topically related to the BDI-II symptoms.

To search for symptoms of depression in Reddit data, we proposed an approach based on semantic search using as queries the corresponding responses for each item from BDI-II. Synthetic data provided by LLMs have been successfully used in other works [141, 146, 148, 331] and have proved to be a reliable method for augmenting and fine-tuning downstream models. Inspired by these approaches, we experimented with generating synthetic Reddit posts using `text-davinci-3` [150] and Llama-3 8B[6] for each item of the BDI-II questionnaire. We designed a prompt such that the answers had more diversity than the BDI-II responses and conveyed the intimate way of sharing experiences and feelings specific to Reddit [280]. In Table 4.9, we showcase the prompt we used to instruct the model, similar to the approach of Wang et al. [331]. However, our prompt was simpler, and geared towards simulating user posts. We included instructions that limited the size of the text, ensured semantic diversity in the generated texts and ensured that the generated data contained emotional and anecdotal experiences that aligned with each BDI-II item. In Algorithm 1, we show our algorithm for generating data using the OpenAI text completion API[7]. Each completion was post-processed by removing trailing quotation marks, enumeration numbers, and splitting by *newline* to obtain individual texts. We used the same prompt to generate synthetic data with Llama-3 8B. BDI-II contains 21 items related to depression symptoms, with a total of 90 possible responses measuring the intensity of symptoms. For each of these 90 responses, we generated 30 synthetic Reddit posts, totaling 2,700 generated texts by GPT-3.5 and 2,700 generated texts by Llama-3. We show in Table 4.10 some GPT-3.5 generated examples for the first symptom of the BDI-II questionnaire. In Table 4.11, we show Llama-3 generated examples for the second symptom of the BDI-II questionnaire. The generated texts were longer than the BDI-II responses and had greater diversity. Some examples even contained self-disclosure, which is specific for Reddit data [280], such as "My cat passed away" and "I just broke up with my partner". We hypothesized that, by augmenting the queries with the synthetically generated data, we would be able to retrieve more relevant sentences.

We pre-processed all the posts from each Reddit user by removing URLs and texts not in English, detected by the *polyglot*[8] package. We computed the sentence embeddings for all texts (original posts, BDI-II responses, generated data) using two state-of-the-art methods, a variant of MPNet [332] and MentalRoBERTa [135]. The MPNet variant we used, `multi-qa-mpnet-base-dot-v1`[9] was explicitly designed for semantic search. MentalRoBERTa[10] was chosen because it was trained on a

[6] `https://huggingface.co/meta-llama/Meta-Llama-3-8B`
[7] `https://platform.openai.com/docs/guides/gpt/completions-api`
[8] https://github.com/aboSamoor/polyglot/
[9] `https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1`
[10] `https://huggingface.co/mental/mental-roberta-base`

```
You are asked to come up with a set of "{N}" diverse reddit posts that are
examples to the BDI depression questionnaire for the "{symptom}" symptom. For
this symptom, the BDI answer of interest is "{item}".
These examples will be given to a ranking model that will compute the similarity
between the answer item text and the reddit post.

Here are the requirements:
1. The language used for the reddit posts should be diverse. For example, you
should combine descriptions of past experiences with feelings or events.
2. The reddit posts should be in English.
3. The reddit posts should be 2 to 3 sentences long.
4. The reddit posts should provide substantial content to make ranking feasible.
5. The reddit posts should be specific and not just describe general situations,
but rather specific personal experiences and self-disclosure.
6. The reddit posts should, as much as possible, not contain the
exact words of the BDI item.

List of "{N}" reddit posts:
```

Table 4.9: Instructions for the data generation that ensure diversity and adherence to the BDI-II responses.

**Algorithm 1** Algorithm for generating synthetic data based on BDI-II items and responses.

**Require:** instruction prompt $P$, questionnaire $Q$
  $D \leftarrow \{\}$
  **for** symptom $s$ and item $q \in Q$ **do**
    $p = P$.format(N = $N$, item = $q$, symptom = $s$)
    $d$ = `openai.Completion.create`(model = `"text-davinci-003"`, prompt = $p$)
    $d$ = `postprocess`($d$)
    $D = D \cup \{\text{d}\}$
  **end for**
    **return** $D$

large corpus of mental health data, mainly from Reddit. The MentalRoBERTa model has shown good downstream performance for mental disorders detection from social media data [333, 334].

We performed semantic search and used cosine similarity between embeddings to get the most relevant social media posts to the original BDI-II responses and generated queries. We retrieved the top 50 sentences with the highest cosine similarity to the queries. Given that we could submit a maximum of 1,000 results for each submission, we sorted the retrieved sentences by cosine similarity scores and kept only the most relevant 1,000 sentences.

We experimented with different queries and embedding methods. The official submissions of our BLUE team are detailed below:

**SemSearchOnBDI2QueriesMPNet** We performed semantic search using the original 90 BDI-II responses as queries. All texts were encoded using MPNet.

**SemSearchOnGPTGeneratedQueriesMPNet** We performed semantic search

Table 4.10: Examples generated by GPT 3.5 for the first symptom of BDI-II: Sadness.

| BDI-II Response | Generated text |
|---|---|
| I do not feel sad. | My cat passed away last week and although I miss her terribly, I'm not feeling any sadness about it.<br>I just broke up with my partner, but I'm focusing on all the good memories we shared and not letting myself be consumed by sadness. |
| I feel sad much of the time. | I feel like I'm stuck in an endless cycle of sadness and emptiness.<br>It feels like no matter how hard I try, I can't shake this sadness that's been following me around. |
| I am sad all the time. | I'm feeling so empty and down lately, like nothing can make me happy anymore.<br>I feel so overwhelmed with sadness that it's hard to get out of bed in the morning. |
| I am so sad or unhappy that I can't stand it. | I feel like I'm stuck in a dark hole with no way out and it's suffocating me.<br>I'm so overwhelmed by sadness that I can barely function anymore. |

Table 4.11: Examples generated by LLama-3 for the second symptom of BDI-II: Pessimism.

| BDI-II Response | Generated text |
|---|---|
| I am not discouraged about my future. | Although life is unpredictable, I'm hopeful that I'll be able to find joy in whatever comes my way.<br>No matter how difficult things get, I believe in myself and my ability to turn any situation around. |
| I feel more discouraged about my future than I used to be. | I'm feeling increasingly pessimistic about my prospects and it's taking a toll on my mental health.<br>All my plans seem to have gone up in smoke and I'm starting to doubt whether I'll ever reach my goals. |
| I do not expect things to work out for me. | I know I shouldn't even bother dreaming about a better future because it'll never happen.<br>I'm so used to failure that I don't even expect success anymore. It's just not an option. |
| I feel my future is hopeless and will only get worse. | I can't help but think that all my struggles now are leading up to an even more difficult time ahead.<br>Every day I wake up with this heavy weight on my chest, knowing that things are only going to get worse. |

using the 2,700 GPT-3.5 generated synthetic Reddit texts as queries, with MPNet embeddings.

**SemSearchOnGPTGenerated&BDI2QueriesMPNet** We use all the original and GPT-3.5 generated queries and perform semantic search on texts encoded with MPNet.

**SemSearchOnBDI2QueriesMentalRoberta** We use the original BDI-II responses as queries, but MentalRoBERTa is used for embedding the data.

**SemSearchOnGPTGeneratedQueriesMentalRoberta** We perform semantic search using the GPT-3.5 generated data as queries; texts were encoded using MentalRoBERTa.

In addition to our official submissions, we tested two other approaches using Llama-3 generated queries:
**SemSearchOnLlamaGeneratedQueriesMPNet** We performed semantic search using the Lama-3 generated synthetic texts as queries, with MPNet embeddings.
**SemSearchOnLlamaGeneratedQueriesMentalRoberta** We perform semantic search using the Llama-3 generated data as queries; texts were encoded using MentalRoBERTa.

### 4.3.2 Results

Table 4.12: Ranking-based evaluation (majority voting) for Task 1: Search for symptoms of depression. *Denotes that the submission was not officially submitted to the eRisk Lab. **BLUE** represents the name of our team.

| Team | Run | AP | R-PREC | P@10 | NDCG@1000 |
|---|---|---|---|---|---|
| Formula-ML | SentenceTransformers_0.25 | **0.319** | **0.375** | **0.861** | **0.596** |
| OBSER-MENH | salida-distilroberta-90-cos | 0.294 | 0.359 | 0.814 | 0.578 |
| uOttawa | USESim | 0.160 | 0.248 | 0.600 | 0.382 |
| NailP | T1_M2 | 0.095 | 0.146 | 0.519 | 0.226 |
| RELAI | bm25\|mpnetbase | 0.048 | 0.081 | 0.538 | 0.140 |
| UNSL | Prompting-Classifier | 0.036 | 0.090 | 0.229 | 0.180 |
| UMU | LexiconMultilingualSentenceTransformer | 0.073 | 0.140 | 0.495 | 0.222 |
| GMU | FAST-DCMN-COS-INJECT_FULL | 0.001 | 0.003 | 0.014 | 0.005 |
| Mason-NLP | MentalBert | 0.035 | 0.072 | 0.286 | 0.117 |
| **BLUE** | SemSearchOnBDI2QueriesMPNet | 0.104 | 0.126 | 0.781 | 0.211 |
| **BLUE** | SemSearchOnGPTGenerated&BDI2QueriesMPNet | 0.065 | 0.086 | 0.629 | 0.160 |
| **BLUE** | SemSearchOnGPTGeneratedQueriesMPNet | 0.052 | 0.074 | 0.586 | 0.139 |
| **BLUE** | SemSearchOnBDI2QueriesMentalRoberta | 0.027 | 0.044 | 0.386 | 0.089 |
| **BLUE** | SemSearchOnGPTGeneratedQueriesMentalRoberta | 0.029 | 0.063 | 0.367 | 0.105 |
| **BLUE*** | SemSearchOnLlamaGeneratedQueriesMPNet | 0.007 | 0.023 | 0.142 | 0.005 |
| **BLUE*** | SemSearchOnLlamaGeneratedQueriesMentalRoberta | 0.002 | 0.008 | 0.076 | 0.002 |

Table 4.13: Ranking-based evaluation (unanimity) for Task 1: Search for symptoms of depression. *Denotes that the submission was not officially submitted to the eRisk Lab.

| Team | Run | AP | R-PREC | P@10 | NDCG@1000 |
|---|---|---|---|---|---|
| Formula-ML | SentenceTransformers_0.25 | 0.268 | **0.360** | **0.709** | **0.615** |
| Formula-ML | SentenceTransformers_0.1 | **0.293** | 0.350 | 0.685 | 0.611 |
| OBSER-MENH | salida-distilroberta-90-cos | 0.281 | 0.344 | 0.652 | 0.604 |
| uOttawa | USESim | 0.139 | 0.232 | 0.438 | 0.380 |
| NailP | T1_M2 | 0.090 | 0.143 | 0.410 | 0.229 |
| UMU | LexiconMultilingualSentenceTransformer | 0.059 | 0.125 | 0.333 | 0.209 |
| RELAI | bm25\|mpnetbase | 0.039 | 0.069 | 0.343 | 0.124 |
| UNSL | Prompting-Classifier | 0.020 | 0.063 | 0.090 | 0.157 |
| GMU | FAST-DCMN-COS-INJECT_FULL | 0.001 | 0.003 | 0.014 | 0.006 |
| Mason-NLP | MentalBert | 0.024 | 0.054 | 0.190 | 0.099 |
| **BLUE** | SemSearchOnBDI2Queries | 0.129 | 0.167 | 0.643 | 0.260 |
| **BLUE** | SemSearchOnAllQueries | 0.067 | 0.105 | 0.452 | 0.177 |
| **BLUE** | SemSearchOnGeneratedQueries | 0.052 | 0.088 | 0.381 | 0.147 |
| **BLUE** | SemSearchOnBDI2QueriesMentalRoberta | 0.032 | 0.058 | 0.300 | 0.104 |
| **BLUE** | SemSearchOnGeneratedQueriesMentalRoberta | 0.018 | 0.059 | 0.186 | 0.085 |
| **BLUE*** | SemSearchOnLlamaGeneratedQueriesMentalRoberta | 0.003 | 0.010 | 0.052 | 0.001 |
| **BLUE*** | SemSearchOnLlamaGeneratedQueriesMPNet | 0.007 | 0.028 | 0.084 | 0.003 |

The results of the eRisk Lab task on searching for depression symptoms are presented in Tables 4.12 and 4.13. We show the results of all five official runs submitted by our team and the best-performing runs from other teams. We also present the results of the two runs using Llama-3 generated data. The metrics used to evaluate the sentences' relevance were Average Precision (AP), R-Precision, P@10, and NDCG@1000. Table 4.12 presents the systems' performance compared to the gold standard obtained from majority voting of the relevant sentences assessed by the annotators. Table 4.13 presents the systems' performance compared to the gold standard obtained from the sentences considered relevant by all three annotators. Comparing our proposed methods, the model using only the BDI-II responses as queries, SemSearchOnBDI2QueriesMPNet, performed best in both ranking-based evaluation settings, majority voting and unanimity, achieving 0.104 AP in the first scenario, and 0.129 AP in the second one. The second-best model was the one that used as queries all the texts (original and generated), SemSearchOnGPTGenerated&BDI2QueriesMPNet, with an AP of 0.065 in majority voting evaluation, and 0.067 in unanimity evaluation. The models using only generated data from GPT-3.5 and Llama-3 as queries had the lowest performance. The SemSearchOnBDI2QueriesMPNet model had a good P@10 of 0.781 for majority voting ranking-based evaluation and 0.643 for unanimity evaluation, showing that our semantic search method using MPNet embeddings on the original BDI-II queries was best at retrieving relevant sentences in top 10 documents. Even if the embeddings provided by the pre-trained model on mental health data, MentalRoBERTa, had a good performance for detection tasks [333, 334], it had the lowest performance for symptoms retrieval.

However, our proposed methods ranked fourth compared to all the systems developed by other participants in the eRisk task. Our hypothesis that the synthetically generated queries will improve performance was proved false. We aimed for variability, as the BDI-II responses were short and standard, but the texts generated by GPT-3.5 and Llama-3 might be too specific. Some of the generated texts provided too many details, which were not helpful for semantic search: e.g., "I just got back from a great vacation and it's been really hard to get back into the swing of things - not feeling particularly sad, but definitely a bit down.", "I don't know what to do with myself anymore - no matter how hard I try, I can't shake this overwhelming sense of gloom.".

In Figure 4.5, we present the most common words in the original BDI-II queries, generated queries with GPT-3.5 and Llama-3 8B, and the manually annotated relevant documents. Llama-3 generated queries are more diverse than the ones generated by GPT-3.5. Even though "feel" is the most common word in all the data, it is noticeable that the BDI-II queries and the relevant documents focus on more specific depression symptoms, such as "appetite", "interested", "tired", "sleep", "energy", "restless". The queries generated by GPT-3.5, on the other hand, contain anecdotal experiences (as shown in Table 4.10), but the most common words in all of them are generic.




Figure 4.5: Comparison of most frequent words between the original BDI-II queries, generated queries and the manually annotated relevant documents.

## 4.4 Summary

In this chapter, we addressed our **RQ3**, by employing the recent advancements in natural language processing in our participation in the eRisk Lab. We have shown that recent models, specifically transformer-based models can detect patterns of mental disorders more effectively than earlier deep learning models. We presented our contribution to three editions of the eRisk Lab on the early detection of mental health problems using social media texts. In the eRisk 2021 Lab, our approaches tackled early risk detection of pathological gambling and self-harm [2]. For early risk detection of problem gambling, we opted for crawling and constructing a dataset from popular subreddits that address gambling addiction and trained a BERT classifier on user posts, with control users from RSDD and eRisk 2018, two popular datasets for depression detection. Similarly, we crawled the subreddit associated with self-harm behavior to augment the existing training dataset provided by the competition organizers for the detection of self-harm. We obtained competitive results through our methods.

For our second participation in the eRisk Lab, we proposed a transformer architecture that performs user-level classification of gambling addiction and

depression detection [7]. For each individual, the transformer processes a set of texts encoded by a pretrained sentence encoder to model the interactions between posts and mitigate noise in the dataset. Our model is interpretable and allows for automatic dataset creation by filtering uninformative posts in a user's history. Our method is a promising approach, especially for social media text processing, where a user has many texts: some informative and some unrelated to the particular modeling task. However, their interaction is indicative of the mental state of the user. We achieved the best $ERDE_5$ score of 0.015, and the second-best $ERDE_{50}$ score of 0.009 for pathological gambling detection. For the early detection of depression, we obtained the second-best $ERDE_{50}$ (0.027).

Lastly, we presented our contributions to the eRisk 2023 Lab task on retrieving relevant social media text relevant to the symptoms of depression from the BDI-II questionnaire [10, 18]. We performed semantic search using the original BDI-II responses and synthetically generated texts as queries. We hypothesized that, by using GPT-3.5 and Llama-3 to generate synthetic data similar to Reddit posts in which users disclose their feelings and experiences, we could retrieve more relevant sentences for each BDI-II item. We experimented with two pretrained transformer-based methods to encode the queries and social media posts, MentalRoBERTa and a variant on MPNet designed specifically for semantic search. Our hypothesis was shown to be false; the model performing semantic search using as queries the original BDI-II responses yielded more relevant sentences than the one using generated data. The synthetic data generated by GPT-3.5 and Llama-3 were too specific for retrieving depression symptoms, and future work needs to be done to generate suitable data for this task.

In conclusion, to answer **RQ3**, we demonstrated in this chapter that computational approaches using recent advancements in natural language processing are more effective at detecting the early risk of mental disorders from social media data compared to earlier methods that relied on traditional machine learning or deep learning models.

CHAPTER 5

# Multimodal Depression Detection from Social Media Data

This chapter presents our proposed computational approaches for depression detection using multimodal data consisting of text, images and videos, based on our published works:

- **A.-M. Bucur**, A. Cosma, P. Rosso and L. P. Dinu, *It's Just a Matter of Time: Detecting Depression With Time-Enriched Multimodal Transformers*, Advances in Information Retrieval. ECIR 2023. Lecture Notes in Computer Science 13980 (2023) pp. 200–215
- D. Gimeno-Gómez, **A.-M. Bucur**, A. Cosma, C.-D. Martínez-Hinarejos and P. Rosso, *Reading Between the Frames: Multi-Modal Depression Detection in Videos From Non-Verbal Cues*, Advances in Information Retrieval. ECIR 2024. Lecture Notes in Computer Science 14608 (2024) pp. 191–209

Although unimodal models have demonstrated good performance in detecting depression from social media text [2, 49, 50], multimodal models have been shown to outperform them [51–53]. We propose to evaluate the effectiveness of multimodal models in two different contexts. We first develop a time-enriched multimodal transformer to detect depression from social media text and images. In addition, we develop a temporal model that processes various non-verbal cues across time to estimate depression from video content.

In this chapter, we aim to respond to the following research question:

**RQ4** Are multimodal models more effective than unimodal models in detecting depression from social media data that includes text, images, or video content?

## 5.1 Depression Detection using Textual and Visual Information

Detecting depression through user-generated content on the internet has been a longstanding area of interest within the research community, offering valuable tools for the screening of mental health issues. The widespread use of social media platforms provides an ideal opportunity to explore how mental health issues are expressed in posts and interactions among users. Current methods for depression detection from social media mainly focus on text processing, and only a few also use images posted by users. In this work, we propose a flexible time-enriched multimodal transformer architecture for detecting depression from social media posts, using pre-trained models for extracting image and text embeddings. Our model operates directly at the user level, and we enrich it with the relative time between posts by using time2vec positional embeddings. Moreover, we propose another model variant, which can operate on randomly sampled and unordered sets of posts to be more robust to dataset noise. Using EmoBERTa and CLIP embeddings, we show that our method surpasses other methods on two multimodal datasets, obtaining state-of-the-art results of 0.931 F1 score on a popular multimodal Twitter dataset and 0.902 F1 score on the only multimodal Reddit dataset.

To date, many methods have been developed to estimate signs of mental disorders, such as depression and eating disorders, based on users' social media content [74, 182]. The primary focus has been on processing the text of these posts, driven in part by the widespread availability and performance of pre-trained language models like BERT [49, 50, 335]. Textual and visual information have recently been used for multimodal depression detection in social media data. This includes datasets collected from platforms such as Twitter [52, 76, 163], Instagram [51, 128], Reddit [64], Flickr [53], and Sina Weibo [336]. While multimodal approaches have shown promising results, they generally assume that social media posts are made synchronously and at regular intervals.

We propose a time-enriched multimodal transformer for user-level depression detection from social media posts (i.e., posts with text and images). Instead of operating at the token level in a low-level manner, we use the cross- and self-attention mechanism across posts to learn the high-level posting patterns of a particular user. The attention layers process semantic embeddings obtained by pre-trained state-of-the-art language and image processing models. Our method does not require major architectural modifications as opposed to current time-aware methods for mental disorders detection [167, 337]. It can easily accommodate temporal information by simply manipulating the transformer positional encodings (e.g., using time-enriched encodings such as time2vec [338]). We propose two ways to train our architecture: a time-aware approach using time2vec and a set-based, in which we do not employ positional encodings and regard the user posts as a set. The second approach is motivated by the work of [339], which observed that positional encodings are not universally necessary to obtain good downstream performance.

We train and evaluate our method on two social media multimodal datasets, each with its particularities in user posting behavior, and obtain state-of-the-art results in depression detection.

### 5.1.1 Time-Enriched Multimodal Transformers

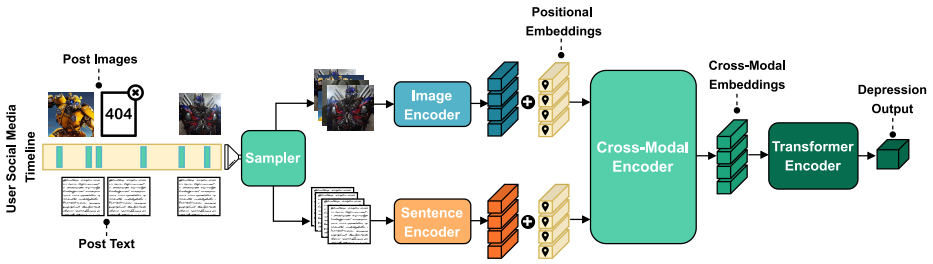



Figure 5.1: The overall architecture of our proposed method for multimodal depression detection from images and texts. We sample several posts containing text and images from a user's social media timeline. Each image and text is encoded with pre-trained image and text encoders. The sequence of encoded images and texts is then processed using a cross-modal encoder followed by a transformer encoder, and the relative posting time is encoded into the positional embeddings. After mean pooling, we perform classification for the particular user.

The problem of depression detection from social media data is usually formulated as follows: given a user with an ordered, asynchronous sequence of multimodal social media posts containing text and images, determine whether or not the user has symptoms of depression. This formulation corresponds to user-level binary classification. The main difficulty in this area is modeling many user posts (i.e., tens of thousands), in which not all posts contain relevant information for depression detection. This aspect makes the classification inherently noisy. We consider a user $i$ to have multiple social media posts $U_i$, each containing the posting date $\Delta$, a text $T$ and an accompanying image $I$. A post-sequence $P_i$ is defined as $K$ posts sampled from $U_i$: $P_i = \{(T^j, I^j, \Delta^j) \sim U_i, j \in (1 \dots K)\}$. During training, we used an input batch defined by the concatenation of $n$ such post-sequences: $B = \{S_{b_1}, S_{b_2}, \dots S_{b_n}\}$. Since the users' posts are asynchronous (i.e., are not regularly spaced in time), time-aware approaches based on T-LSTM [167, 168] have become the norm in modeling users' posts alongside with the relative time between them. However, in T-LSTM [337], including a relative time component involves the addition of new gates and hand-crafted feature engineering of the time. Moreover, T-LSTM networks are slow, cannot be parallelized and do not allow for transfer learning.

To address this problem, we propose a transformer architecture that can perform user-level multimodal classification, as shown in Figure 5.1 (Bucur et al. [13]). To process the multimodal data from posts, we first encode the visual and textual

information using pre-trained models (e.g., CLIP [340] for images and EmoBERTa [341] for text). The embeddings are linearly projected to a fixed size using a learnable feed-forward layer, are augmented with a variant of positional encodings (expanded below) and are further processed with a cross-modal encoder based on LXMERT [342]. Finally, self-attention is used to process the cross-modal embeddings further, and classification is performed after mean pooling. The network is trained using the standard binary cross-entropy loss. In this setting, the transformer attention does not operate on low-level tokens such as word pieces or image patches. Rather, the cross- and self-attention operates across posts, allowing the network to learn high-level posting patterns of the user, and be more robust to individual uninformative posts. As opposed to other related works in which a vector of zeros replaces the embeddings of missing images [163, 164], we take advantage of the attention masking mechanism present in the transformer architecture [324] and mask out the missing images.

For this work, we experiment with positional embeddings based on *time2vec* [338] to make the architecture time-aware. Utilizing *time2vec* as a method to encode the time $\tau$ into a vector representation is a natural way to inject temporal information without any major architectural modification, as it was the case for T-LSTM, for example. *Time2vec* has the advantage of being invariant to time rescaling, avoiding hand-crafted time features, it is periodic, and simple to define and consume by various models. It is a vector of $k + 1$ elements, defined as:

$$\mathbf{t2v}(\tau)[i] = \begin{cases} \omega_i g(\tau) + \phi_i & i = 0 \\ \mathcal{F}(\omega_i g(\tau) + \phi_i) & 1 \leq i \leq k \end{cases} \tag{5.1}$$

where $\mathbf{t2v}(\tau)[\mathrm{i}]$ is the $i^{\text{th}}$ element of $\mathbf{t2v}(\tau)$, $\mathcal{F}$ is a periodic activation function (in our case it is $sin(x)$), and $\omega_i s$ and $\phi_i s$ are learnable parameters. To avoid arbitrarily large $\tau$ values, we transform $\tau$ with $g(\tau) = \frac{1}{(\tau+\epsilon)}$, with $\epsilon = 1$. For processing user's posts, we use sub-sequence sampling and sample $K$ *consecutive* posts from a user's timeline (Figure 5.2). We name the model utilizing *time2vec* embeddings *Time2VecTransformer*.

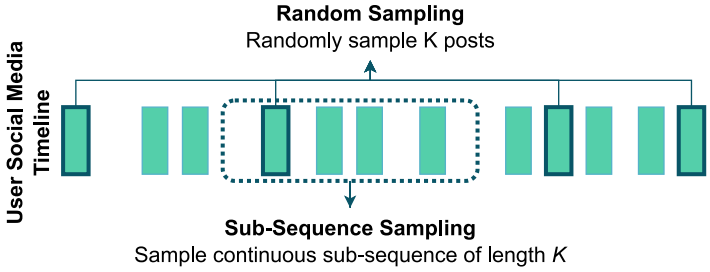



Figure 5.2: The two methods used for sampling posts from a user's timeline. Methods that use positional embeddings (either time2vec or learned) employ sub-sequence sampling, and the SetTransformer, with zero positional embeddings, uses random sampling of posts. $K$ refers to the number of posts in the post window.

However, processing user posts sequentially is often not desirable, as clusters of posts might provide irrelevant information for depression detection. For instance, sub-sequence sampling of posts from users with mental health problems might end up in an interval of positive affect, corresponding to a sudden shift in mood [106]. Somewhat orthogonal to the time-aware school of thought, we also propose a *SetTransformer* for processing sets of user posts for depression detection, to alleviate the issues mentioned above. Our proposed *SetTransformer* randomly samples texts from a user and assumes no order between them by omitting the positional encoding, essentially making the transformer permutation invariant [324]. For this method, K posts in a post-sequence are randomly sampled (Figure 5.2). For *SetTransformer*, we treat the user timeline as a "bag-of-posts" motivated by the work of Dufter, Schmitt and Schütze [339], in which the authors show that treating sentences as "bag-of-words" (i.e., by not utilizing any positional embeddings) results in a marginal loss in performance for transformer architectures.

### 5.1.2 Experiments

To benchmark our method, we use in our experiments *multiRedditDep* [64] and the Twitter multimodal depression dataset from Shen et al. [76] and Gui et al. [163], which were described in Section 2.5. Reddit and Twitter are two of the most popular social media platforms where users choose to disclose their mental health problems [280, 343]. Moreover, the data from these two platforms have different particularities: both social media platforms support images, but the textual information is richer on Reddit, with posts having up to 40,000 characters, as opposed to Twitter, where the limit is 280 characters. In some subreddits from Reddit, the image cannot be accompanied by text; the post is composed only of image(s) and a title with a maximum length of 300 characters. On Twitter, posts with images have the same character limit as regular posts.

Due to the lack of access to the benchmark train/test splits from An et al. [164], we run the experiments following the same experimental settings as Gui et al. [163], and perform five-fold cross-validation. In Table 5.1, we showcase the statistics for both datasets. There is a greater total number of posts from the non-depressed group, but depressed users have more posts, on average, than non-depressed users, in the case of Reddit. The opposite is true for Twitter, with non-depressed users having more posts on average. It is important to note that not all posts have images; only a small number of social media posts contain both text and image data. Figure 5.3 (left) shows that Reddit users have a greater average number of posts than Twitter users. Regarding posting frequency, Twitter users post more frequently than Reddit users, as shown in Figure 5.3 (right), in which the average time between posts is 26.9 hours for Reddit and 11.2 hours for Twitter.

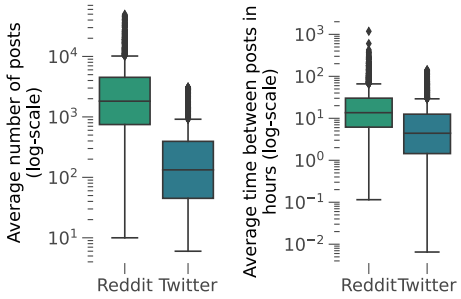



Figure 5.3: Distribution of posts per user and the average time between the posts on Reddit and Twitter multimodal datasets. Left - Distribution of posts per user on both datasets. Users from Reddit have significantly (Mann Whitney U Test, $p < 0.001$) more posts than users from Twitter. Right - Distribution of average time duration in hours between posts at the user level. Users from Twitter post significantly ($p < 0.001$) more frequently than users from Reddit.

| **Dataset** | **Class** | #T | #(T+I) | $\overline{\#T}$ | $\overline{\#(T+I)}$ |
|---|---|---|---|---|---|
| Reddit | Depr | 6,6M | 46.9k | 4.7k | 33.33 |
| | Non-Depr | 8,1M | 73.4k | 3.5k | 31.51 |
| Twitter | Depr | 213k | 19.3k | 152.43 | 13.81 |
| | Non-Depr | 828k | 50.6k | 590.82 | 36.16 |

Table 5.1: Statistics for the Reddit [64] and Twitter [163] datasets. #T represents the number of posts with only text, #(T+I) represents the number of posts with both text (title, in the case of Reddit) and images. $\overline{\#T}$ and $\overline{\#(T+I)}$ represent the average number of posts with only text and text + image per user, respectively.

## Experimental Settings

**Image representation methods.** To encode the images in users' posts, we opted to use two different encoders trained in a self-supervised manner. Many approaches for transfer learning employ a supervised network, usually pre-trained on ImageNet [344]. However, transfer learning with self-supervised trained networks is considered to have more general-purpose embeddings [345], which aid downstream performance, especially when the training and transfer domains qualitatively differ [346]. Since images in our social media datasets are very diverse, including internet memes and screenshots [64] with both visual and textual information, specialized embeddings from a supervised ImageNet model are not appropriate [8]. Therefore, we used **CLIP** [340], a vision transformer trained with a cross-modal contrastive objective, to connect images with textual descriptions. CLIP has been shown to perform well on a wide range of zero-shot classification tasks, and is capable of encoding text present in images. CLIP embeddings are general, multi-purpose, and are trained on a large-scale internet-scraped dataset. Additionally, we used **DINO** [345], a vision transformer pre-trained in a self-supervised way on ImageNet,

achieving notable downstream performance. Moreover, compared to a supervised counterpart, DINO automatically learns class-specific features without explicitly being trained to do so.

**Text representation methods.** We explored three pre-trained transformer models to extract contextual embeddings from users' texts. **RoBERTa** [327], pre-trained in a self-supervised fashion on a large corpus of English text, was chosen given its state-of-the-art performance on downstream tasks. Emotion-informed embeddings from **EmoBERTa** [341] that incorporate linguistic and emotional information from users' posts were also used. Emotions expressed in texts are a core feature used for identifying depression [323, 347], users with depression showing greater negative emotions [40]. EmoBERTa is based on a RoBERTa model trained on two conversational datasets and adapted to identify the emotions found in users' utterances. **Multilingual MiniLM** [328], the distilled version of the multilingual language model XLM-RoBERTa [310], was used for encoding text because both Twitter and Reddit datasets contain posts in various languages besides English, such as Spanish, German, Norwegian, Japanese, Romanian and others. Moreover, MiniLM is a smaller model, providing 384-dimensional embeddings, as opposed to 768-dimensional embeddings from RoBERTa and EmoBERTa.

**Positional encodings.** Since the network is processing sequences of posts, we explored three different methods for encoding the relative order between posts. Firstly, we used a standard **learned** positional encoding, used in many pre-trained transformer models, such as BERT [335], RoBERTa [327] and GPT [150]. Wang and Chen [348] showed that for transformer encoder models, learned positional embeddings encode local position information, which is effective in masked language modeling. However, this type of positional encoding assumes that users' posts are equally spaced in time. Second, we used **time2vec** [338] positional encoding, which allows the network to learn a vectorized representation of the relative time between posts. Lastly, we do not use any positional encoding and treat the sequence of user posts as a set. We refer to these positional encodings as **zero** encodings. For both *learned* and *time2vec*, we used a sub-sequence sampling of posts, while for *zero*, we used a random sample of user posts (Figure 5.2).

### Comparison Models

We evaluate our proposed method's performance against existing multimodal and text-only state-of-the-art models on the two multimodal datasets from Twitter and Reddit, each with its different particularities in user posting behavior. For *multiRedditDep*, since it is a new dataset, there are no public benchmarks besides the results of Uban, Chulvi and Rosso [64]. We report Accuracy, Precision, Recall, F1, and AUC as performance measures. The baselines are as follows. **Time-Aware LSTM (T-LSTM)** [337] - we implement as text-only baseline a widely used [167, 168] T-LSTM-based neural network architecture that integrates the time irregularities of sequential data in the memory unit of the LSTM. **EmoBERTa Transformer** - we train a text-only transformer baseline on user posts. Both text-only models

are based on EmoBERTa embeddings. **LSTM + RL** and **CNN + RL** [349] - two text-only state-of-the-art models which use a reinforcement learning component for selecting the posts indicative of depression. **Multimodal Topic-Enriched Auxiliary Learning (MTAL)** [164] - a model capturing the multimodal topic information, in which two auxiliary tasks accompany the primary task of depression detection on visual and textual topic modeling. **Multimodal Time-Aware Attention Networks (MTAN)** [167] - a multimodal model that uses as input BERT [335] textual features, InceptionResNetV2 [350] visual features, posting time features and incorporates T-LSTM for taking into account the time intervals between posts and self-attention. **GRU + VGG-Net + COMMA** [163] - in which a reinforcement learning component is used for selecting posts with text and images that are indicative of depression and are classified with an MLP. For extracting the textual and visual features, GRU [351] and VGGNet [352] were used. **BERT + word2vec embeddings** - baseline proposed by Uban, Chulvi and Rosso [64], which consists of a neural network architecture that uses as input BERT features from posts' titles and word2vec embeddings for textual information found in images (i.e., ImageNet labels and text extracted from images). **VanillaTransformer** - the multimodal transformer proposed in this work, with standard *learned* positional encodings. **SetTransformer** - the set-based multimodal transformer proposed in this work employing *zero* positional encoding, alongside a random sampling of user posts. **Time2VecTransformer** - the time-aware multimodal transformer proposed in this work using time-enriched positional embeddings (i.e., *time2vec* [338]) and sub-sequence sampling.

#### Training and Evaluation Details

We train all models using Adam [294] optimizer with a base learning rate of 0.00001. The learning rate is modified using Cyclical Learning Rate schedule [353], which linearly varies the learning rate from 0.00001 to 0.0001 and back across 10 epochs. The model has 4 cross-encoder layers with 8 heads each and an embedding size of 128. The self-attention transformer has 2 layers of 8 heads each and the same embedding size.

At test time, since it is unfeasible to use all users' posts for evaluation, with some users having more than 50k posts, we make 10 random samples of post-sequences for a user, and the final decision is taken through majority voting on the decisions for each post-sequence. In this way, the final classification is more robust to dataset noise and uninformative posts.

### 5.1.3 Results

#### Performance Comparison with Prior Works

In Table 5.2, we present the results for models trained on the Twitter dataset. For our models and proposed baselines, each model was evaluated 10 times, and we reported mean and standard deviation. Our architecture, *Time2VecTransformer*, achieves

state-of-the-art performance in multimodal depression detection, obtaining a 0.931 F1 score. The model uses textual embeddings extracted from EmoBERTa and visual embeddings extracted from CLIP as input. The best model uses sequential posts as input from a 512 window size, and the time component is modeled by *time2vec* positional embeddings. Our time-aware multimodal architecture surpasses other models such as T-LSTM [337] and MTAN [167]. Moreover, our method surpasses text-only methods such as T-LSTM and EmoBERTa Transformer - an unimodal variant of our model using only self-attention on text embeddings.

Table 5.2: Results for multimodal depression detection on Twitter dataset [163]. Our models use EmoBERTa and CLIP for extracting embeddings. VanillaTransformer and SetTransformer were trained on 128 sampled posts, while Time2VecTransformer was trained on a window size of 512 posts. Denoted with bold are the best results for each column. *An et al. [164] report the performance on a private train/dev/test split, not on five-fold cross-validation. ** Cheng and Chen [167] are not explicit in their experimental settings for the Twitter data. † indicates that the result is a statistically significant improvement over SetTransformer ($p < 0.005$, using Wilcoxon signed-rank test). ‡ indicates that there is a statistically significant improvement over Time2VecTransformer ($p < 0.05$, using Wilcoxon signed-rank test).

| **Method** | **Modality** | **F1** | **Prec.** | **Recall** | **Acc.** |
|---|---|---|---|---|---|
| T-LSTM [337] | T | $0.848_{\pm 8e\text{-}3}$ | $0.896_{\pm 2e\text{-}2}$ | $0.804_{\pm 1e\text{-}2}$ | $0.855_{\pm 5e\text{-}3}$ |
| EmoBERTa Transformer | T | $0.864_{\pm 1e\text{-}2}$ | $0.843_{\pm 1e\text{-}2}$ | $0.887_{\pm 3e\text{-}2}$ | $0.861_{\pm 1e\text{-}2}$ |
| LSTM + RL [349] | T | 0.871 | 0.872 | 0.870 | 0.870 |
| CNN + RL [349] | T | 0.871 | 0.871 | 0.871 | 0.871 |
| MTAL [164]* | T+I | 0.842 | 0.842 | 0.842 | 0.842 |
| GRU + VGG-Net + COMMA [163] | T+I | 0.900 | 0.900 | 0.901 | 0.900 |
| MTAN [167]** | T+I | 0.908 | 0.885 | 0.931 | - |
| Vanilla Transformer **(ours)** | T+I | $0.886_{\pm 1e\text{-}2}$ | $0.868_{\pm 2e\text{-}2}$ | $0.905_{\pm 2e\text{-}2}$ | $0.883_{\pm 5e\text{-}3}$ |
| SetTransformer **(ours)** | T+I | $0.927_{\pm 8e\text{-}3}$ | $0.921_{\pm 1e\text{-}2}$ | $\mathbf{0.934_{\pm 2e\text{-}2}}^{\ddagger}$ | $0.926_{\pm 8e\text{-}3}$ |
| Time2VecTransformer **(ours)** | T+I | $\mathbf{0.931_{\pm 4e\text{-}3}}^{\dagger}$ | $\mathbf{0.931_{\pm 2e\text{-}2}}^{\dagger}$ | $0.931_{\pm 1e\text{-}2}$ | $\mathbf{0.931_{\pm 4e\text{-}3}}^{\dagger}$ |

Table 5.3: Results for multimodal depression detection on multiRedditDep. Our models use EmoBERTa and CLIP for extracting embeddings. VanillaTransformer was trained on 128 sampled posts, while Time2VecTransformer and SetTransformer were trained on a window size of 512 posts. We denote with bold the best results for each column. *Uban, Chulvi and Rosso [64] conducted experiments using the visual and textual features from images and titles of the posts. † indicates that the result is a statistically significant improvement over Time2VecTransformer ($p < 0.005$, using Wilcoxon signed-rank test).

| **Method** | **Modality** | **F1** | **Prec.** | **Recall** | **Acc.** | **AUC** |
|---|---|---|---|---|---|---|
| T-LSTM [337] | T | $0.831_{\pm 1e\text{-}2}$ | $0.825_{\pm 8e\text{-}3}$ | $0.837_{\pm 1e\text{-}2}$ | $0.872_{\pm 7e\text{-}3}$ | $0.946_{\pm 2e\text{-}3}$ |
| EmoBERTa Transformer | T | $0.843_{\pm 6e\text{-}3}$ | $0.828_{\pm 3e\text{-}3}$ | $0.858_{\pm 1e\text{-}2}$ | $0.879_{\pm 4e\text{-}3}$ | $0.952_{\pm 2e\text{-}3}$ |
| Uban et al. [64]* | T+I | - | - | - | 0.663 | 0.693 |
| VanillaTransformer **(ours)** | T+I | $0.837_{\pm 8e\text{-}3}$ | $0.827_{\pm 1e\text{-}2}$ | $0.848_{\pm 1e\text{-}2}$ | $0.876_{\pm 6e\text{-}3}$ | $0.956_{\pm 3e\text{-}3}$ |
| SetTransformer **(ours)** | T+I | $\mathbf{0.902_{\pm 7e\text{-}3}}^{\dagger}$ | $\mathbf{0.878_{\pm 6e\text{-}3}}^{\dagger}$ | $\mathbf{0.928_{\pm 1e\text{-}2}}^{\dagger}$ | $\mathbf{0.924_{\pm 5e\text{-}3}}^{\dagger}$ | $\mathbf{0.976_{\pm 1e\text{-}3}}^{\dagger}$ |
| Time2VecTransformer **(ours)** | T+I | $0.869_{\pm 7e\text{-}3}$ | $0.869_{\pm 7e\text{-}3}$ | $0.869_{\pm 8e\text{-}3}$ | $0.901_{\pm 5e\text{-}3}$ | $0.967_{\pm 1e\text{-}3}$ |

As opposed to the previous result on Twitter data, which includes the time component, for Reddit, we achieved good performance by regarding the users' posts as a set. In Table 5.3, we showcase our model's performance on the Reddit dataset, compared to Uban, Chulvi and Rosso [64], our trained text-only T-LSTM and text-only EmoBERTa Transformer. Our *SetTransformer* model (with *zero* positional encodings), using as input EmoBERTa and CLIP, obtains the best performance, a 0.902 F1 score. While counter-intuitive, treating posts as a "bag-of-posts" outperforms *Time2VecTransformer* in the case of Reddit. However, the Reddit dataset contains a considerable amount of noise and uninformative posts (links, short comments, etc.), which dilutes discriminative information for depression detection. A random sampling of posts seems to alleviate this problem to some degree.

### Ablation Study

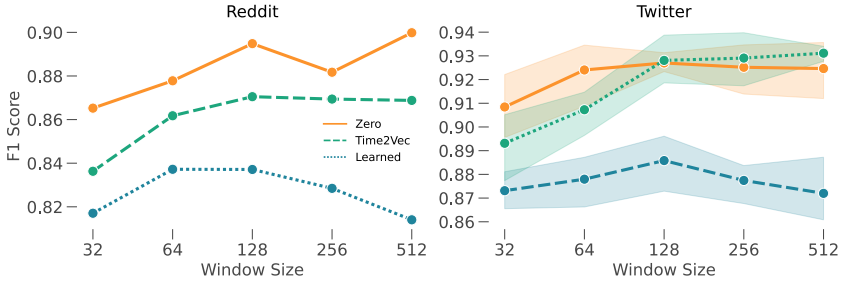



Figure 5.4: Comparison among different window sizes used in our experiments, using EmoBERTa for text embeddings and CLIP for image encodings. For Twitter, the results were averaged across the 5 folds.

To gauge the effect of the sampling window size, we performed experiments using CLIP as an image encoder and EmoBERTa as a text encoder, in which we varied the window size in $K = \{32, 64, 128, 256, 512\}$, as presented in Figure 5.4. For Reddit, the 128 window size best suits all three kinds of positional embeddings, as evidenced by the high F1 score. Even if Reddit users have an average of over 3,000 posts (Table 5.1), 128 posts contain enough information to make a correct decision. On Twitter, for *learned* and *zero* positional embeddings, the model with 128 posts window size performs best, while for *time2vec*, 512 has the best F1 score. This may be because the 512 window size covers the average number of posts from users in the Twitter dataset (see Table 5.1). Given the short period of one month for the posts coming from Twitter, we can hypothesize that for datasets in which the time between posts is very short (average of 11.2 hours for Twitter), the time component modeled by *time2vec* positional encodings may be more informative than other positional embedding methods. For Reddit, many of the users' posts are, in fact, comments or links, which are usually not informative to the model decision. Nevertheless, we achieve a performance comparable to the previous state-of-the-art,

even in low-resource settings, by processing only 32 posts. Including the time component modeled by *time2vec* has a more important contribution when the time between posts is shorter (as in the data from Twitter) as opposed to larger periods between the posts (as is the case of Reddit).

Table 5.4: Model comparison with different text and image encoding methods using the Reddit and Twitter datasets, with time2vec positional embeddings and 128 window size. The best results are with bold, and with underline the second best results.

| | Reddit | | | Twitter | | |
|---|---|---|---|---|---|---|
| **Text+Image Enc.** | **F1** | **Prec.** | **Recall** | **F1** | **Prec.** | **Recall** |
| MiniLM+CLIP | 0.789$_{\pm 6e\text{-}3}$ | 0.686$_{\pm 7e\text{-}3}$ | **0.929$_{\pm 9e\text{-}3}$** | 0.827$_{\pm 8e\text{-}3}$ | 0.803$_{\pm 3e\text{-}2}$ | 0.854$_{\pm 4e\text{-}2}$ |
| MiniLM+DINO | 0.799$_{\pm 9e\text{-}3}$ | 0.745$_{\pm 8e\text{-}3}$ | 0.862$_{\pm 1e\text{-}2}$ | 0.792$_{\pm 8e\text{-}3}$ | 0.782$_{\pm 3e\text{-}2}$ | 0.806$_{\pm 4e\text{-}2}$ |
| RoBERTa+CLIP | 0.845$_{\pm 8e\text{-}3}$ | 0.829$_{\pm 8e\text{-}3}$ | 0.862$_{\pm 1e\text{-}2}$ | **0.943$_{\pm 6e\text{-}3}$** | **0.951$_{\pm 1e\text{-}2}$** | **0.936$_{\pm 2e\text{-}2}$** |
| RoBERTa+DINO | 0.840$_{\pm 9e\text{-}3}$ | 0.820$_{\pm 7e\text{-}3}$ | 0.861$_{\pm 1e\text{-}2}$ | 0.936$_{\pm 1e\text{-}2}$ | 0.946$_{\pm 2e\text{-}2}$ | 0.926$_{\pm 2e\text{-}2}$ |
| EmoBERTa+CLIP | **0.871$_{\pm 7e\text{-}3}$** | **0.883$_{\pm 8e\text{-}3}$** | 0.858$_{\pm 8e\text{-}3}$ | 0.928$_{\pm 1e\text{-}2}$ | 0.933$_{\pm 1e\text{-}2}$ | 0.924$_{\pm 2e\text{-}2}$ |
| EmoBERTa+DINO | 0.863$_{\pm 6e\text{-}3}$ | 0.865$_{\pm 4e\text{-}3}$ | 0.862$_{\pm 1e\text{-}2}$ | 0.915$_{\pm 1e\text{-}2}$ | 0.918$_{\pm 2e\text{-}2}$ | 0.913$_{\pm 2e\text{-}2}$ |

In Table 5.4, we showcase different combinations of text encoders (RoBERTa / EmoBERTa / MiniLM) and image encoders (CLIP / DINO). The difference between image encoders is marginal, due to the small number of images in the users' timeline (Table 5.1). Interestingly, RoBERTa embeddings are more appropriate for Twitter, while EmoBERTa is better suited for modeling posts from Reddit. We hypothesize that this is due to the pseudo-anonymity offered through Reddit, encouraging users to post more intimate and emotional texts [280]. Using RoBERTa text embeddings and CLIP image embeddings, we obtain an F1 score of 0.943, which is even higher than the state-of-the-art. Further, the performance of MiniLM for text embeddings is lacking compared to other encoders.

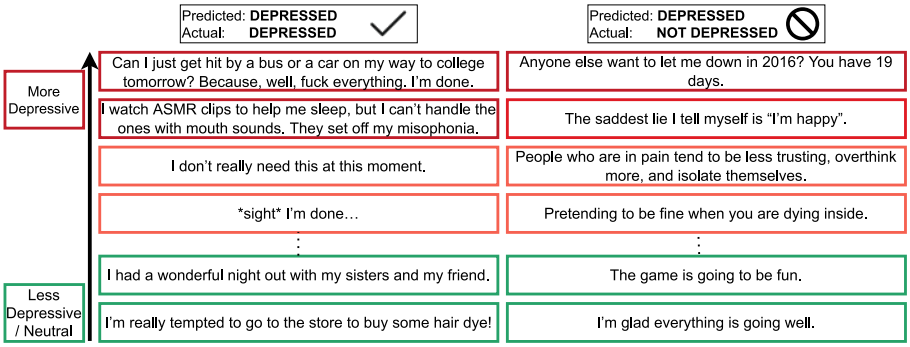



Figure 5.5: Error analysis on two predictions, one correct (Left), and another incorrect (Right). The posts are sorted by their attribution scores given by Integrated Gradients [316]. The top posts have strong attribution for a positive (depressed) prediction, the bottom texts have a weak attribution to a positive prediction. All examples were paraphrased, and only the texts were shown to maintain anonymity.

### Error Analysis

We perform an error analysis on the predictions of *Time2VecTransformer* on the Twitter data. We use Integrated Gradients [316] to extract posts' attributions scores for predictions. In Figure 5.5 (*Left*), the posts with depression cues have the strongest attribution scores, and the model correctly labels the user. Given how the mental health datasets are annotated by users' mention of diagnosis, some users from the non-depressed class may also have depression, but without mentioning it on social media. This may be the case of the user from Figure 5.5 (*Right*), who is showing definite signs of sadness and was incorrectly predicted by the model as having depression. Since our model operates across posts, using a feature attribution method such as Integrated Gradients naturally enables the automatic selection of the most relevant posts from a user, similar to Gui et al. [163] and Ríssola et al. [198], but without relying on a specialized procedure to do so.

## 5.2 Depression Detection using Non-Verbal Cues from Videos

Although research on multimodal depression detection from social media data has been performed using textual, visual, and online behavioral data, visual data is mainly limited to images [13, 48, 54, 163, 216, 222]. However, many social media sites (YouTube, TikTok, Instagram) focus on user-generated video content and, currently, depression detection from in-the-wild video remains largely unexplored. The video modality poses an additional set of challenges for depression detection as it is significantly noisier, video length is highly variable (from shorts of a few seconds to half-hour vlogs in which the person talks directly to the camera), the face and body are not always visible, context rapidly changes, and audio quality is low, especially in amateur videos. Uploaded videos are mostly out of context and stand-alone instead of interactive discussions in social media threads. In this context, the task is to process subtle depression cues from the subject's behavior in the video.

According to the DSM-V criteria [30], depression is manifested through symptoms such as depressed mood, sleep disturbance, appetite changes, loss of interest, etc. In addition, psychomotor changes (agitation or retardation) are a central feature of depression, intertwined with other symptoms, such as loss of energy, fatigue, and lack of concentration [65]. Some of the psychomotor particularities associated with depression are: reduced facial expressiveness [56], slower speech rate, longer pause time [57], and downward head tilt [58]. In addition, studying psychomotor changes in depression is essential because they can predict clinical response to medication in depression [57, 73]. However, these cues have mainly been studied in laboratory settings [268, 270] and significantly less so in in-the-wild videos. To the best of our knowledge, the only available benchmark dataset for depression detection using data from online platforms (i.e., YouTube) is the D-Vlog dataset [177].

In this section, we present a multimodal temporal model that processes multiple non-verbal cues across time to estimate depression from both noisy, in-the-wild videos [177] and from controlled laboratory recordings [268, 270]. Our method is simple and flexible: by using high-level modality extractors, positional embeddings, and modality conditioning vectors, this approach can be easily scaled to an arbitrary number of modalities.

### 5.2.1 Multimodal Transformer Architecture for Video Content

Inspired by psychological research [56, 58, 68, 71, 72], we employ additional features related to motor manifestations of depression, such as face and audio embeddings, hand and body landmarks, blinking, and gaze patterns. Our relatively simple architecture can process multiple modalities across different frame rates and processes only relevant parts of the video to perform classification. In contrast to previous approaches that mainly use global information [174, 177, 178], we design

our architecture to handle both local, frame-level information, as well as temporal video dynamics.

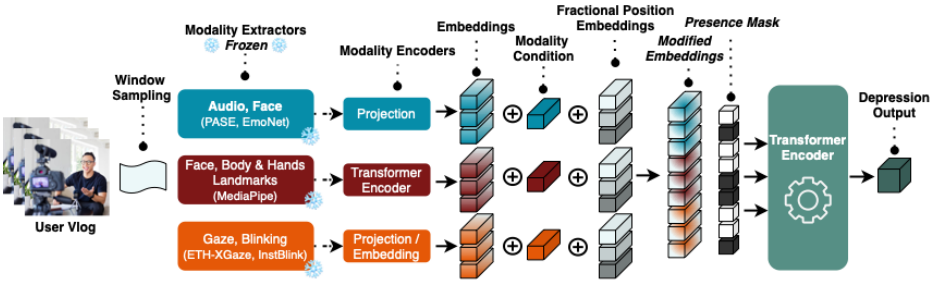



Figure 5.6: The overall architecture of our proposed method for depression detection using non-verbal cues. We extract high-level non-verbal cues using pretrained models, process them using a modality-specific encoder, condition the resulting embeddings with positional and modality embeddings, and process the sequence with a transformer encoder to perform the final classification.

## Model Architecture

**Overview** Given a dataset of vlogs with people directly talking to the camera, our aim is to train a model to estimate whether the subject in a video has depression or not. In this formulation, depression detection amounts to video-level classification. However, performing the classification of a high-level psychological affliction directly from pixels is infeasible without impractical amounts of labeled data. Consequently, in our work, we considered only high-level non-verbal cues, and discarded any textual information to remove any conversational topic bias [83]. The way most depression datasets are collected is based on direct mention of diagnosis [74, 354] and by using only non-verbal cues, the model's performance is a more accurate depiction of its performance in realistic settings. Non-verbal cues are extracted using state-of-the-art pretrained models [355–360] and subsequently processed to perform classification. Figure 5.6 provides a high-level overview of our pipeline for depression classification from video. We further present a way to properly handle videos of different framerates and video lengths, and process an arbitrary amount of modalities.

Formally, given a dataset of depression labeled videos $\mathcal{D} = \{(V_1, \hat{y}_1), \dots (V_n, \hat{y}_n)\}$, where $V_i$ is the video and $\hat{y}_i$ its corresponding label, the goal is to find the optimal parameters $\theta$ of a model $f_\theta$ that minimize the average cross entropy loss across the dataset:

$\hat{\theta} = \arg\min_\theta \mathbb{E}_{1 \le k \le n}[\mathcal{L}(f_\theta(V_k), \hat{y}_k)]$.

**Window Sampling** Since videos and vlogs in the wild have vastly different durations and framerates (see Fig 5.8), with some videos exceeding 30 minutes, it is computationally unfeasible to process them directly in their entirety. Consequently, we operate on randomly sampled temporal windows $W_l^k \sim V_k$ from each video, with

$W_l^k$ of fixed time in seconds, and $l$ a random temporal index in the video. Temporal windows can cover a different number of total frames, corresponding to each video's framerate. The audio is sampled at 100 frames per second [361], but video framerates from D-Vlog [177] vary between 6 and 30 frames per second. Through the window sampling approach, we assume that the mental health information in the video is time-invariant.

**Modality Extraction and Encoding** For each video, we considered multiple semantic modalities $\mathcal{M} = \{m_1, m_2 \dots\}$, that are extracted using a frozen pretrained model from both audio and visual information of the video at each frame. Consequently, the total number of frames for a video across all temporally concatenated modalities is the sum $T^k = T_1^k + \cdots + T_{|\mathcal{M}|}^k$. Since each modality has a different dimensionality $d_{m_j}$, we uniformize each modality output with an associated learnable modality encoder $E_{m_j} : \mathbb{R}^{d_{m_j}} \rightarrow \mathbb{R}^d$, operating at the frame-level, such that each modality has the same dimensionality $d$. Thus, the feature vector for a window $W_l^k$ and a modality $m_j$ is $x_{m_j} = E_{m_j}(W_l^k)$, with $x_{m_j} \in \mathbb{R}^{T_j \times d}$. Details about each modality encoder $E_{m_j}$ are showcased in Section 5.2.1. Furthermore, each modality extractor can signal the presence of a modality in a frame, which allows us to construct a binary presence mask $M_j^k \in \mathbb{R}^{T_j^k}$. The presence mask is later used in attention masking for the final transformer encoder [324]. In practice, for in-the-wild videos, we first extracted and cached all modalities for each video, and sampled windows such that the subject is present in at least 50% of the frames in a window. This is because some videos were amateur-made and extremely noisy, and the subject's body, hands and face were not always visible; for instance, some people talk while driving their car and the steering wheel obscures the camera, or the subject walks out of the scene.

**Modality Condition** We distinguish between the embeddings of different modalities by using an additional learned condition embedding $e_{m_j} \in \mathbb{R}^{|\mathcal{M}| \times d}$ that is added to each vector of its corresponding modality. This is similar to how BERT-type models [362] differentiate between sentences in tasks such as natural language inference.

**Fractional Position Embeddings** Due to the mismatch between sampling rates, we temporally align the frames of acoustic and video-based modalities by using fractional positional embeddings [363] of the form $p_{m_j} \in \mathbb{R}^{\max_i(T_i) \times d}$. For a video, we construct a matrix of positional embeddings of size $\max_i(T_i)$, corresponding to the modality with the highest sample rate, usually audio. Afterwards, positions for a modality $m_j$ are then uniformly indexed according to the ratio $r = \lfloor \max_i(T_i)/T_{m_j} \rfloor$. Fractional positional embeddings are based on sinusoidal positional embeddings [324] and are not learned during training. Frames across modalities have the same positional embedding at the same corresponding positions in time, irrespective of the sampling rate (see Figure 5.7). Readers are referred to Harzig, Einfalt and Lienhart [363] for more details.

An individual modified feature vector becomes $\bar{x}_{m_j} = x_{m_j} + e_{m_j} + p_{m_j}$. Finally, we process the concatenated modified feature vectors $X_l^k = \bar{x}_{m_1} \parallel \bar{x}_{m_2} \parallel \dots \parallel$

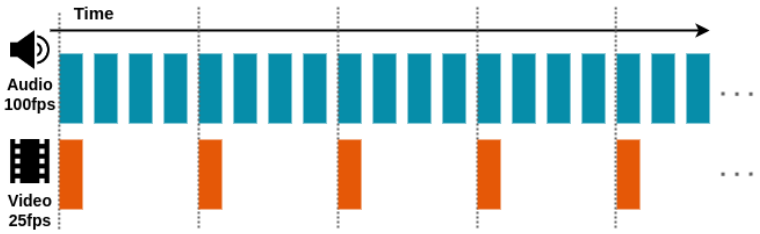



Figure 5.7: Illustrative example of fractional positional embedding for temporally aligning video and audio sampling rates, similar to the work of Harzig, Einfalt and Lienhart [363].

$\bar{x}_{|\mathcal{M}|}$ alongside each corresponding presence mask $M_l^k = M_1 \parallel M_2 \parallel \ldots \parallel M_{|\mathcal{M}|}$ from each modality of a sampled temporal window with a transformer encoder network [324] to obtain a prediction $y = f_\theta(X_l^k, M_l^k)$, with $X_l^k \in \mathbb{R}^{T_l^k \times |\mathcal{M}| \times d}$ and $M_k \in \mathbb{R}^{T_l^k \times |\mathcal{M}|}$. The final optimization formulation can be described as $\hat{\theta} = \arg\min_\theta \mathbb{E}_{(1 \leq k \leq n, 1 \leq l \leq |V_k|)}[\mathcal{L}(f_\theta(X_l^k, M_l^k), \hat{y}_k)]$.

**Discussion** In our framework, modality fusion is done immediately after encoding each modality with the learnable encoders $E_{m_j}$, and the predominant part of the computation is performed with the main transformer encoder. This type of fusion can be considered "early" fusion [364]. Furthermore, as defined by Vaswani et al. [324], the scaled dot-product attention is permutation invariant across the sequence without adding positional embeddings [348]. However, we employ two types of positional embeddings: learned modality conditions and fractional positional embeddings. In a single timestep $t$, the modality vectors $x_{m_j}^t$ are permutation invariant, sharing the same positional embedding. The order of the modalities does not affect the final computation, as the modalities are considered an unordered set, corresponding to the SetTransformer [326] formulation. Other approaches [174, 177, 178] mainly perform a global computation across a coarse temporal sampling of modalities, which is prone to lose the information of the finer non-verbal cues in manifestations of depression. However, through our window sampling, modality conditions, and positional embeddings, the transformer attention operates locally, inter-modality (at the same timestep), and globally, intra-modality (across multiple timesteps), to perform a successful depression classification. In this setting, the model performs a soft cross-attention between modalities. Furthermore, compared to other approaches in this area [174–176], which propose overly-complicated methods with modest results, our approach is simple, flexible, easy to reproduce, and it surpasses other methods on the three main datasets we evaluated.

### Extracting and Encoding Non-Verbal Cues

**Psychological Motivation** Psychomotor manifestations are one of the main features of depression [65] found in the DSM-V criteria and self-assessment tools, such as PHQ-9 or BDI-II. The non-verbal facial cues associated with depression are reduced expressiveness [56], and fewer smiles [66]. Regarding body language, a downward

head tilt [58], longer gait cycles, and a slower gait cadence [67] are common indicators. Regarding hand movements, individuals with depression may use fewer gestures [68] and exhibit impaired gesture performance [69]. A slower speech rate characterizes the voice, with longer pauses between words [57], and a decrease in loudness [70]. The eyes undergo subtle shifts with an increased blinking rate [71], reduced eye contact, and increased gaze angle [72]. Given these non-verbal changes in depressed individuals, we investigated how they interact and complement each other in depression detection. This work refers to major depressive disorder (MDD), not other related mood disorders, such as bipolar disorder, peripartum disorder, etc.

**Audio-Visual Embeddings (AV)** We extract semantic embeddings for the audio channels and the face gestures using pretrained models. For **audio embeddings**, we used the Pyannote Audio toolkit[1] [365, 366] to detect voice activities and extract 256-dimensional voice embeddings using the PASE+ model [355, 367] in the time slots in which the person is speaking. The PASE+ model provides general embeddings [367, 368] as it was trained in a task-agnostic manner in noisy and reverberant environments. For extracting emotion-informed **face embeddings**, we used EmoNet[2] [356], a model capable of estimating both discrete emotion classes and their corresponding valence and arousal [369, 370] measures. We obtain a 256-dimensional embedding vector by taking the embedding from the last layer before the classification of EmoNet. For both types of embeddings, in our model, we process the embeddings with a 1D batch normalization layer followed by a linear layer to project the embeddings into a 256-dimensional space.

**Landmarks (LM)** We extract relevant facial, body, and hands keypoints using pretrained networks. For **face landmarks**, we used the Face Alignment Network model [357] to extract 68 2-dimensional facial landmarks alongside their corresponding confidence score. For **body landmarks**, we used the MediaPipe[3] [358, 371] toolkit to extract 33 3-dimensional landmarks. For **hand landmarks**, we used the hand landmark detector from the MediaPipe [358] toolkit to extract 21 3-dimensional landmarks for each hand. Our model processes the landmarks using a 1D batch-normalization layer followed by a 2-layer transformer encoder that operates on the spatial dimension, obtaining the final embedding through average spatial pooling. Additionally, for encoding hand landmarks, we distinguish between right and left hands by adding an embedding vector, similar to the modality condition described above.

**Eye Information (EYES)** We extract gaze and blinking information as additional modalities. For **gaze tracking** we used the official implementation[4] of the ETH-XGaze gaze estimator [359] to extract the gaze direction represented by 3 angle coordinates. For **blinking patterns**, we used the official implementation of the

[1] https://github.com/pyannote/pyannote-audio.
[2] https://github.com/face-analysis/emonet.
[3] https://github.com/google/mediapipe.
[4] https://github.com/hysts/pytorch_mpiigaze_demo.

InstBlink model[5] [360], and labeled the video frames where the person is blinking (0 for not blinking and 1 for blinking). To encode gaze, we process the angles using a 1D batch normalization followed by a linear layer to project the angle vector into a 256-dimensional embedding. To encode blinking, we used one of two different learnable 256-dimensional embeddings at each frame, corresponding to blinking or not blinking states.

### 5.2.2 Experiments

Benchmark datasets for studying the non-verbal behavior of subjects with depression include primarily data collected in laboratory settings, e.g., the DAIC-WOZ [268], E-DAIC [270], Depression Severity Interviews Database [372], Audio-Visual Depressive Language Corpus [373], and the BlackDog Database [374]. These datasets consist of clinical interviews or footage with subjects performing different tasks (reading certain texts, counting, etc.). However, research in depression detection from video in general, in-the-wild, scenarios is limited. To the best of our knowledge, D-Vlog [177] is the only dataset containing in-the-wild user-generated videos. In our experiments, we use D-vlog [177], an in-the-wild dataset sourced from YouTube, and two datasets collected in laboratory settings, DAIC-WOZ [268] and Extended DAIC-WOZ [270]. The three datasets were presented in Section 2.5.

In our work, in addition to studying the original audio-visual features provided by the authors of D-vlog, we explored the multiple non-verbal cues described in Subsection 5.2.1. When extracting additional features, some of the videos were unavailable (deleted or made private by the owner), resulting in 861 videos. Specifically, 70, 8, and 23 videos were missing for training, validation, and test, respectively.

For DAIC-WOZ, in addition to the transcript of the entire interview, COVAREP audio- [375] and OpenFace video-based [376] features are provided. Original recordings are not available for privacy reasons, and we used the original features provided by the authors. We used 7 modalities: the acoustic COVAREP features, the five vocal tract resonance frequencies, the 68 3-dimensional facial landmarks, the facial action units [170], the gaze tracking, and the head pose landmarks. We did not use the text transcripts in our model. In our DAIC-WOZ experiment, we consider only the PHQ-8 binary labels. These binary labels are derived from the PHQ-8 score from each subject, with those who scored 10 or higher being labeled as depressed and the rest belonging to the control group.

In addition to the interview transcript, multiple audio-visual features are provided for Extended DAIC-WOZ, based on both expert knowledge-based methods [377, 378] and deep learning representations [379]. Similar to DAIC-WOZ, we only used the features provided by the authors. Concretely, we used a total of 6 modalities: the 13 mel frequency cepstral coefficients [377] alongside their first and second derivatives, the 88 eGeMAPS measures [378] (covering information related to

[5] https://github.com/wenzhengzeng/MPEblink.

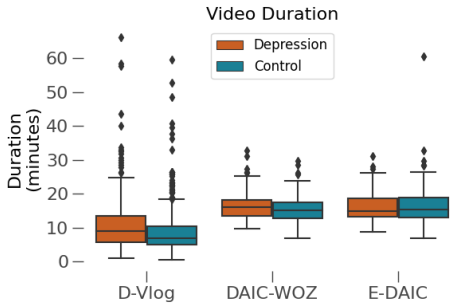



Figure 5.8: Distributions of video durations for each of our benchmarking datasets: D-Vlod, DAIC-WOZ and E-DAIC.

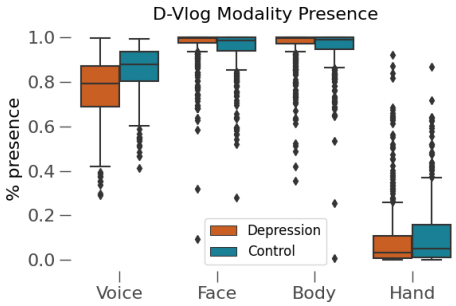



Figure 5.9: Presence distributions for each of our considered modalities in D-Vlog.

spectral, prosodic, and voice quality patterns), face embeddings extracted by a pre-trained ResNet-50 [379], and the OpenFace features [376], including gaze tracking, head pose landmarks and facial action units. We did not use the text transcripts in our model. Similar to DAIC-WOZ, we use the PHQ-8 binary labels as ground truth.

We showcase the video durations for the datasets in Figure 5.8. DAIC-WOZ and E-DAIC videos have a longer overall duration (15.9 and 16.2 minutes, respectively) than D-Vlog and the video lengths vary considerably less. Notably, in D-Vlog, the depression group's videos are longer than those of the control group, with averages of 10.5 and 8.9 minutes, respectively. In Figure 5.9, we show the percentage of presence for each modality of interest in the D-Vlog dataset. Hands are the modality that is the least present in the vlogs. Even if the face and body are present in most of the videos, there are still many outliers in which the modalities are present less than 80% of the time. As expected, the voice is also commonly present in the videos, but less in the depression group. The fact that in-the-wild vlogs may contain different transitions or sequences without the subject's presence in the video motivated us to incorporate the presence mask in our architecture.

### Evaluation

Unlike other approaches, which either did not specify a formal evaluation protocol [176] or processed the features truncated to the average length of videos [175, 177], we perform sequential evaluation across all temporal windows. For each non-overlapping window in a video $\{W_1, W_2, \ldots W_n\} \in V_k$ taken in order (containing at least 50% of modalities for D-Vlog), we perform depression classification and obtain a sequence of predictions: $\{y_1, y_2, \ldots y_n\}$, with $y_l = f_{\hat{\theta}}(W_l)$. The final decision is performed by voting on the decisions across windows: $y = \text{argmax}_{y_i} P(X = y_i)$. This evaluation protocol can be used for an early warning detection system [380] for suicide prevention by taking the decision at an earlier window $n' < n$. For instance, this procedure can be used in live, continuous video streams on platforms such as Facebook or Twitch, in which cases of live suicide have been unfortunately registered in the past [381].

### Implementation Details

Experiments were conducted on 2 GeForce RTX 3060 GPUs with 12GB memory. In all our experiments, we used the AdamW optimizer [382] with a learning rate of 0.001 decaying through a cosine scheduler across 200 epochs, with a batch size of 8. On average, a training run took around 2 hours. In initial hyper-parameter search experiments, we found a suitable configuration for each model. Depending on the number of modalities used, our model's parameters range between 8.4M to 15M. For D-Vlog, we used a context window size of 9 seconds and discarded window samples in which less than 50% of the frames did not contain the subject's face. For DAIC-WOZ and E-DAIC, we did not discard any window, and used a context window of 6 seconds. For D-Vlog and E-DAIC, the transformer encoder was composed of 8 layers with 8 32-dimensional attention heads, while for DAIC-WOZ, we used eight layers with 4 64-dimensional attention heads. Similar to D-Vlog (see Subsection 5.2.1), for encoding the modalities in DAIC-WOZ and E-DAIC we defined a 2-layer transformer encoder when dealing with landmarks, and a linear projection when processing the other non-verbal features. For evaluation, we report mean and standard deviation across five training runs.

### Comparison Methods

**D-Vlog** We compared our proposed method to several state-of-the-art approaches for depression detection from video. In contrast with previous works, which use only low-level facial and audio descriptors [173–178], our model can also incorporate additional high-level facial and audio embeddings, gaze and blinking features, and hands and body landmarks. We also compare our method, which handles both local frame-level information and temporal video dynamics, to previous works that use global information from videos [174, 177, 178], and with previous time-aware methods [173, 175, 176]. We included two strong state-of-the-art methods [173,

176]. Zheng, Yan and Wang [176] proposed a temporal convolutional transformer that uses specialized medical and depression knowledge graphs. Tao et al. [173] proposed a spatio-temporal transformer architecture inspired by [383, 384].

**DAIC-WOZ** We compared our method to other works that approached the task from a non-verbal perspective, without considering the text transcripts. To show the capability of our proposed multimodal architecture, we not only compared our method to those works that used hand-crafted features as our case [385–387], but also to those that explored deep-learning embedded representations [388]. Furthermore, although Wei et al. [387] introduced attention-based modules, most of these works were mainly based on CNNs combined with LSTMs [386–388].

**E-DAIC** Although prior works [176, 389–391] have been developed for this dataset, to the best of our knowledge, none of them explored the binary depression detection task from a non-verbal perspective, as all methods used the text transcript in their models. Consequently, we report the first results in this setting for E-DAIC.

Unlike our method, most previous works on D-Vlog only considered two modalities. Consequently, we introduce an additional strong baseline for general multimodal classification for all three datasets. We used the Perceiver [392] based on an open-source implementation[6], as a general multimodal encoder. However, different from our approach, the Perceiver is based on iterative attention between modalities on a lower-dimensional hidden latent sequence, which can suffer from gradient explosion [393, 394] and loss of information. We constructed the Perceiver to be similar to our model regarding the number of parameters and hidden dimensions.

### 5.2.3 Results

Table 5.5: Results on the D-Vlog test set compared to existing approaches. Best results are highlighted in bold and second best results are underlined.

| Method | Original Split | Additional Modalities | $F_1$ | Precision | Recall |
|---|---|---|---|---|---|
| Yoon et al. [177] | ✓ | ✗ | 0.64 | 0.65 | 0.66 |
| Nguyen et al. [178] | ✓ | ✗ | 0.64 | 0.66 | 0.64 |
| Zheng et al. [176] | ✓ | ✗ | 0.65 | 0.65 | 0.65 |
| Zhou et al. [175] | ✓ | ✗ | 0.66 | 0.66 | 0.67 |
| Zhou et al. [174] | ✓ | ✗ | 0.67 | 0.67 | 0.67 |
| Tao et al. [173] | ✓ | ✗ | <u>0.75</u> | **0.73** | 0.78 |
| Perceiver [392] | ✓ | ✗ | 0.74±0.01 | 0.67±0.03 | <u>0.84±0.07</u> |
| **Ours** | ✓ | ✗ | **0.76±0.01** | <u>0.67±0.02</u> | **0.87±0.02** |
| Perceiver [392] | ✗ | ✗ | 0.72±0.01 | <u>0.66±0.03</u> | 0.80±0.03 |
| Perceiver [392] | ✗ | ✓ | 0.73±0.01 | 0.68±0.08 | 0.81±0.15 |
| **Ours** | ✗ | ✗ | <u>0.73±0.01</u> | 0.62±0.01 | **0.89±0.05** |
| **Ours** | ✗ | ✓ | **0.78±0.01**‡ | **0.74±0.05** | <u>0.84±0.06</u> |

[6] https://github.com/lucidrains/perceiver-pytorch

Table 5.6: Performance of our multimodal method on the DAIC-WOZ test set. Best results are highlighted in bold and second best results are underlined. We show performance across audio-only (A), visual-only (V) and audio-visual (AV) modalities.

| Method | Modality | $F_1$ | Precision | Recall |
|---|---|---|---|---|
| Song et al. [386] | A | 0.46 | 0.32 | 0.86 |
| Ma et al. [388] | A | 0.52 | 0.35 | **1.00** |
| Williamson et al. [385] | A | 0.57 | - | - |
| Wei et al. [387] | A | **0.61** | 0.56 | 0.66 |
| Perceiver [392] | A | 0.33±0.26 | 0.30±0.28 | 0.45±0.21 |
| **Ours** | A | 0.49±0.05 | **0.58±0.05** | 0.47±0.05 |
| Song et al. [386] | V | 0.50 | 0.60 | 0.43 |
| Williamson et al. [385] | V | 0.53 | - | - |
| Wei et al. [387] | V | 0.61 | 0.64 | 0.58 |
| Perceiver [392] | V | **0.62±0.05** | **0.67±0.08** | **0.62±0.05** |
| **Ours** | V | 0.62±0.06 | 0.66±0.04 | 0.61±0.06 |
| Song et al. [386] | AV | 0.50 | 0.60 | 0.43 |
| Wei et al. [387] | AV | 0.61 | **0.78** | 0.50 |
| Perceiver [392] | AV | 0.58±0.13 | 0.75±0.05 | 0.59±0.11 |
| **Ours** | AV | **0.67±0.05** | 0.68±0.04 | **0.66±0.06** |

Table 5.7: Performance of our multimodal method on D-Vlog using different modality combinations, compared to the Perceiver. Best results are highlighted in bold and second best results are underlined. Our method is surpassing the strong multi-modal baseline, the Perceiver.

| Method | Modality | $F_1$ | Precision | Recall |
|---|---|---|---|---|
| Perceiver [392] | AV | 0.69±0.06 | 0.72±0.07 | 0.69±0.13 |
| | AV+EYES | 0.73±0.01 | 0.68±0.08 | 0.81±0.15 |
| | AV+LM | 0.67±0.13 | 0.70±0.14 | 0.74±0.28 |
| | AV+LM+EYES | 0.64±0.17 | **0.79±0.07** | 0.58±0.22 |
| **Ours** | AV | 0.72±0.05 | 0.73±0.03 | 0.72±0.11 |
| | AV+EYES | 0.75±0.04 | 0.72±0.03 | 0.79±0.12 |
| | AV+LM | 0.75±0.03 | 0.73±0.05 | 0.77± 0.06 |
| | AV+LM+EYES | **0.78±0.01** | 0.74±0.05 | **0.84±0.06** |

**D-Vlog** In Table 5.5, we present the results of our method, compared to previous state-of-the-art approaches. Since the release of the D-Vlog dataset, some YouTube videos became unavailable, reducing the test split. To provide a fair comparison, we evaluate our model on both the original test split and the test split of currently available videos from which we extracted additional modalities. Our model achieves a 0.76 $F_1$ using the original low-level facial and audio features provided by D-Vlog authors from the original test split. However, when we evaluate our model using the original features on the reduced test split, the performance of the model is

Table 5.8: Performance of our multimodal method on the E-DAIC test set. Best results are highlighted in bold and second best results are underlined. We show performance across audio-only (A), visual-only (V) and audio-visual (AV) modalities.

| Method | Modality | $F_1$ | Precision | Recall |
|---|---|---|---|---|
| Perceiver [392] | A | 0.36±0.17 | <u>0.66±0.12</u> | 0.41±0.09 |
| | V | 0.39±0.22 | 0.50±0.26 | 0.48±0.18 |
| | AV | <u>0.53±0.09</u> | 0.54±0.05 | <u>0.54±0.11</u> |
| **Ours** | A | 0.51±0.11 | **0.71±0.05** | 0.52±0.08 |
| | V | 0.50±0.08 | 0.58±0.05 | 0.49±0.07 |
| | AV | **0.56±0.12** | 0.59±0.07 | **0.58±0.13** |

lower, with only a 0.73 $F_1$, pointing to the fact that the samples from the reduced test set are more challenging to classify than the original test split. Using the additional modalities (audio-visual, landmarks and eyes) inspired by psychological research, our model obtains the best performance of 0.78 $F_1$, surpassing the previous state-of-the-art method proposed by Tao et al. [173]. Moreover, compared to the Perceiver, a state-of-the-art method for multimodal classification, our model achieves a considerable improvement of 5% $F_1$.

We provide the results of our method using different modalities in Table 5.7. Even if the Perceiver model is a strong baseline for multimodal information processing, our proposed method surpasses it in all settings. Adding **LM** (face, body and hands landmarks) and **EYES** features (gaze and blinking) separately offers a slight improvement over the **AV** setting using only face and audio embeddings. Our model obtains the best performance with all modalities **AV+LM+EYES**, achieving an $F_1$ of 0.78. Results show that different non-verbal manifestations of depression are needed for accurate detection from video data. Our best-performing model uses emotionally informed face embeddings from EmoNet, audio embeddings, spatial information from the face, body, and hand landmarks, and gaze and blinking patterns.

**DAIC-WOZ** Even though the DAIC-WOZ dataset was extensively used for depression detection, most previous works incorporate text-based features extracted from the video transcripts, with state-of-the-art approaches achieving performances of around 0.96 $F_1$ [176]. However, this high performance is not due to a deep understanding of depression, but is rather due to spurious correlations between the text transcript and the depression label, as the dataset contains recordings of clinical interviews with predefined questions from PHQ-8. For this reason, we are using only non-verbal cues and discarding the text transcript to provide a more realistic performance measurement for depression detection. The methods we considered in Table 5.6 do not incorporate textual features for depression modeling. Our model is competitive with other methods in video-only settings, but we achieve the best performance of 0.67 $F_1$ using both modalities. Our method performs better in audio and audio-visual settings than the Perceiver baseline.

**E-DAIC** Similar to DAIC-WOZ, the current state of the art for binary classification explored text-based features extracted from the transcripts, a considerably easier formulation than only using non-verbal cues, achieving high performance [176]. In Table 5.8, we report the first results addressing the task from a non-verbal perspective on binary classification for E-DAIC. Our model substantially improves when combining both audio and visual modalities and outperforms the Perceiver in all scenarios, showing that our architecture is highly efficient at integrating multiple modalities.

**Explainability** Any model developed to work with data from humans, especially in clinical scenarios, needs to be interpretable and explainable, to increase trust and transparency, ethical and legal compliance, and to allow researchers and engineers to debug and perform error analysis. Consequently, we present a potential way of explaining multimodal depression detection models from video using Integrated Gradients [316]. In Figure 5.10, we show attribution scores for six modalities in a selected window from a video from D-Vlog, with higher values corresponding to a strong attribution towards a positive prediction. We show the attribution values across time, for the video frames (bottom) and the audio frames (top), sampled at a higher frequency. In this particular context window, movement through the body and hand landmarks and gaze are more indicative of the mental state of the subject, compared to audio or facial expressions. From around frame 150, the person is not visible in the video (while audio is still present), and consequently, the attribution scores for each missing modality are 0. This shows that our model can properly handle missing modalities. Our analysis provides the first steps in automatic interpretation and understanding of how different non-verbal cues contribute to the manifestations of depression.

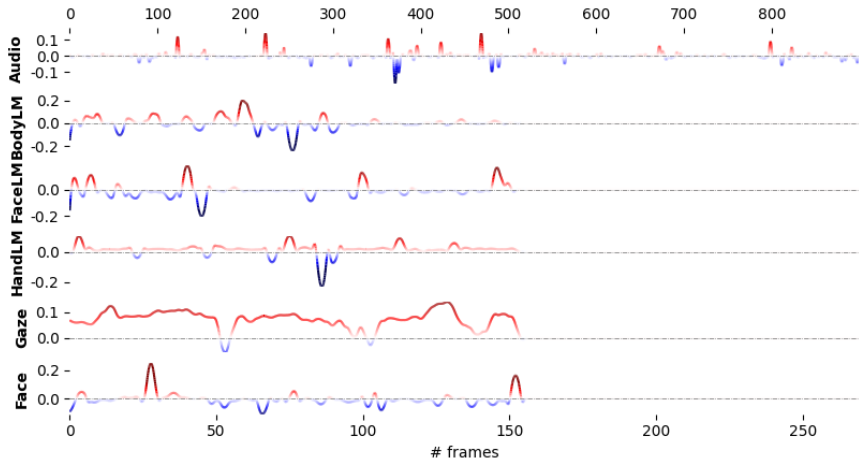



Figure 5.10: Attribution scores per each modality across frames obtained with Integrated Gradients [316] on a selected window from a subject suffering from depression from D-Vlog. Higher values correspond to a strong attribution towards a positive prediction.

## 5.3 Summary

In this chapter, we addressed **RQ4** and presented two different multimodal architectures that can handle textual and visual information from social media posts (Twitter and Reddit) or non-verbal cues from YouTube videos. Our model for depression detection from social media posts [13] is designed to operate directly at the user level: the attention mechanisms (both cross-modal and self-attention) attend to text and images across posts, not individual tokens. We provided two viable ways to train our method: a time-aware approach, which encodes the relative time between posts through *time2vec* positional embeddings, and a set-based approach, in which no order is assumed between users' posts. We experimented with multiple sampling methods and positional encodings (i.e., *time2vec*, *zero* and *learned*) and multiple state-of-the-art pretrained text and image encoders. Our proposed *Time2VecTransformer* model achieves state-of-the-art results, obtaining a 0.931 average F1 score on the Twitter depression dataset [163]. Using the *SetTransformer*, we get a 0.902 F1 score on *multiRedditDep* [64], a multimodal depression dataset with users from Reddit. Given the particularities of the two datasets, we hypothesize that a set-based training regime is better suited to handle datasets containing large amounts of noise and uninformative posts. At the same time, a time-aware approach is suitable when user posting frequency is high.

In addition, we presented a simple and flexible multimodal transformer architecture capable of detecting depression from multiple non-verbal cues from video (Gimeno-Gómez, Bucur et al. [14]). Following psychological studies [56–58, 65], we explored additional high-level modalities and showed an improvement in depression detection from in-the-wild videos. We obtained state-of-the-art results in three key benchmarking datasets for depression detection, both in-the-wild vlogs [177] and in recorded clinical interviews [268, 270], surpassing previous works [173–178, 385–388] by a considerable margin. Finally, we showed that our model is interpretable, and it can provide importance scores of each modality of a particular subject across time, making our method a viable solution for depression estimation, potentially in early-warning detection and suicide prevention in platforms with continuous video streams.

CHAPTER 6

# The Expression of Happiness in Social Media of Individuals Reporting Depression

This chapter presents an exploration of happy moments in social media texts from individuals mentioning a depression diagnosis compared to a control group, based on our published papers:

- **A.-M. Bucur**, A. Cosma and L. P. Dinu, *Life Is Not Always Depressing: Exploring the Happy Moments of People Diagnosed With Depression*, Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC) (2022) pp. 4186–4192
- **A.-M. Bucur**, B. Chulvi, A. Cosma and P. Rosso, *The Expression of Happiness in Social Media of Individuals Reporting Depression*, IEEE Transactions on Affective Computing 16.1 (2025) pp. 360–375

Depression has long been studied in the NLP field, with most works focusing on individuals' negative emotions. People with depression experience happiness, but this has not been extensively studied. In this chapter, we conduct a large-scale study of happy moments from social media texts of individuals mentioning a depression diagnosis. We develop a computational framework to extract happy moments from text, and annotate them with semantic topics, gender labels, and agency and sociality measures. We aim for our research to be used to inform psychology interventions, which can foster feelings of longer-lasting happiness and represent a promising path of collaboration between computational linguistics and psychology.

We seek to answer the following research question:

**RQ5** How can deep learning models identify happy moments from the social media data of individuals with self-reported depression diagnoses? Are there differences in these happy moments compared to a control group?

In this chapter, we focus on people's expressions of momentary happiness, namely happy moments [395], in social media. Pleasure (hedonic enjoyment) and self-satisfaction (personal expressiveness) are the two major conceptions of happiness in philosophy [396]. Consequently, a happy moment represents the natural language expression of an experience that brings feelings of pleasure or self-satisfaction to an individual. Previous works [395, 397] exploring the happy moments from user-generated content have shown that happiness is not expressed in natural language using only positive words (e.g., "exciting", "successfully", "pleasure") in contexts such as "I was offered an exciting professional opportunity" or "I was able to successfully change the engine air filter in my car" [398]. It can also contain words with negative polarity (e.g., "illness", "die", "pain", "worry") in contexts such as: "Being allowed to bring my grandfather home on hospice. Since I can take care of him, he gets to die in his own home not a cold hospital." or "when my dog didn't die from an illness". Furthermore, the happy moments reported by people might not contain any positive or negative words, and be expressed in a neutral manner: "Having a beer after work.", "I mopped the kitchen.", "I got a full 8 hours of sleep." This suggests that the expression of happiness is more than just disclosing an experience using positive or negative-loaded words.

According to Jaidka et al. [395], the current state-of-the-art methods for detecting emotions do not work properly in identifying happy moments. Similarly, Moreno-Ortiz et al. [397] showed that sentiment classification models are ineffective in detecting positive sentiment in all happy moments. Therefore, these methods are not suitable for detecting expressions of happiness. Since there is no existing approach that focuses on detecting happy moments, we propose the development of automatic models that can learn from an existing corpus of happy moments [398] using a positive-unlabeled approach [399]. To evaluate the performance of our pipeline for happy moments detection, we use a subset of happy moments from social media that have been manually annotated.

Based on the Agency-Communion framework [400], the authors from Jaidka et al. [395] propose that the expression of happiness is related to agency and sociality. The authors use the term "sociality"[1] as an equivalent to "communion", which includes not only social interactions with other people but also social identity. Agency means that the author is in control of the happy moment they are experiencing (e.g., "I successfully took care of a large problem at work.", "I just bought a car."), while sociality implies that there are other people involved in the happy moment, besides the author (e.g., "I was happy when my daughter and I made dinner together.", "My niece and I were having fun playing a video game together.") [401]. While agency and sociality have been explored before in happy moments from the general population, to the best of our knowledge, we are the first to explore the two concepts in writings from individuals who have a depression diagnosis. The dataset that is currently available with annotations for agency and sociality in happy moments,

[1] In our empirical work, we also use the term "sociality", following a similar approach to Jaidka et al. [395].

as described in Jaidka et al. [401], only has 1/6 of its data annotated, while the remaining 5/6 of the data is unlabeled happy moments. Due to this limitation, developing models for this task is challenging and requires a semi-supervised approach. Consequently, we develop a semi-supervised classification model for detecting agency and sociality that is on par with the previous state-of-the-art approaches.

Scholars from psychology have largely studied the relationship between the expression of happiness and depression. Life satisfaction and happiness are negatively associated with depression [59, 60]. However, the lives of people diagnosed with depression are not always depressing; they also experience happiness. Bergsma et al. [61] measured the happiness of people with and without a mental disorder (including depression) and showed that 68.4% of individuals with psychiatric disorders reported that they often felt happy, in comparison with 89.1% of control individuals. Spinhoven et al. [402] compared the happiness of people having a depression diagnosis with individuals having a co-morbid social anxiety disorder, and a control group (without any of the two disorders). Over 90% of the people from the control group reported that they were happy, while for the depression group, the percentage was approx. 60%. In addition, less than 30% of the individuals from the co-morbid group reported that they were happy.

In line with these results revealing that individuals with depression experience happiness, we aim to extract the expression of happiness in written text coming from individuals mentioning a depression diagnosis online to understand the experiences that contribute to their happiness.

For a better presentation of this chapter, we divide the main research question introduced at the beginning of the chapter into four sub-questions:

**RQ5 (a)** How can deep learning models extract happy moments from the social media data of individuals with depression?

**RQ5 (b)** To what extent can the sense of agency and sociality be inferred from free-form text?

**RQ5 (c)** What are the topics of happy moments of individuals with mentions of a depression diagnosis compared with the ones from the control group? Are these topics gender-dependent?

**RQ5 (d)** What role do sociality and agency play in the expression of happy moments for men and women mentioning a depression diagnosis compared with the control group?

To answer our research questions, we develop an extensive computational framework that automatically extracts relevant information for our study using a combination of five different existing datasets. We first extract happy moments from social media text using positive-unlabeled learning, and filter them using contextual emotions. We annotate each happy moment with agency and sociality measures

with custom-trained models. Further, we extract the topics for each happy moment in our dataset using a semantic topic modeling method. Finally, we follow the methodology from Tigunova et al. [403] to label each user from our dataset with gender information (Bucur et al. [19]).

## 6.1 Depression and Expressions of Happiness

The corpus-based exploration of happiness shared online began with the work of Mihalcea and Liu [404], which explored data from LiveJournal[2] and developed a recipe for happiness. The authors concluded that certain ingredients contribute to people's happiness: novelty, food, drinks and a social context (being in the company of others).

Other approaches for understanding how people express their happiness in text rely on crowd-sourced data. Asai et al. [398] developed HappyDB, a large-scale dataset with 100,000 happy moments from Mechanical Turk workers. The crowd workers were asked to answer two questions: "What made you happy in the past 24 hours?" and "What made you happy in the past 3 months?". The purpose of the dataset was to understand the activities that induce a state of happiness and who is involved in these activities. To enrich HappyDB, Jaidka et al. [395] further collected agency and sociality information for the happy moments of crowd-sourced workers following the Agency-Communion framework [400]. Annotators were asked if, for a given text, the author of the described happy moment is or is not in control of the action (agency) and if the moment is in a social context (sociality). The two datasets were used in the CL-Aff Happiness Shared Task [401], aiming to predict agency and sociality and extract happiness insights. Moreno-Ortiz et al. [397] performed a keyword analysis on HappyDB to identify the sources of happiness, and their results suggest that there are two types of happiness: internal and external. The external sources of happiness come from social activities and interactions, but also reference places, events, or hobbies. The internal sources of happiness come from actions that people do individually, or are more personal. The authors also report that an important source of happiness is represented by consumerism and materialism, which come in different forms, such as shopping, traveling, or having new experiences.

The works mentioned above have focused on the happiness expressed by the general population, but happiness, or the absence of it, plays a vital role in mental disorders [61, 402]. The automatic detection of mental health problems from social media data covers a wide range of disorders, such as depression [13, 347, 405], self-harm [74, 120], suicide ideation [406], eating disorders [323], pathological gambling [2, 317] and post-traumatic stress disorder [407]. Even if the field of depression detection from user-generating content has seen tremendous advancements in the last few years [13, 405, 408], the works that focus on the emotions of depressed

[2] https://www.livejournal.com/.

individuals [154, 323, 347] do not pay closer attention to the happiness expressed by people with depression. In the current work, we do not aim to perform depression detection but rather study the data from social media users that were already annotated with depression labels from online mentions of diagnoses [74].

In a pilot study on the expressions of happiness, we showed that individuals with depression report more meaningful happy moments (involving their friends and family), while people in the control group report happiness about everyday activities (about money and entertainment) (Bucur et al. [9]). In our previous work, we conducted a qualitative exploration of how users with depression express happiness on social media. In this chapter, we take the first step to explore the personal agency and sociality in happy moments coming from individuals with depression. The two concepts are critical in the expressions of happiness [395]. Furthermore, we incorporate the prevalence of emotions and topic modeling in our analysis and consider the larger context in which the happy moment appears. We conduct our experiments on RSDD [74], a large-scale dataset with almost two orders of magnitude more users than the dataset used in our preliminary work [9]. We also use a manually annotated subset of happy moments extracted from social media to evaluate the performance of our pipeline for happy moments detection.

## 6.2 Agency and Communion

Agency and communion were terms developed by Bakan [409] to reflect two fundamental modalities of human actions. While agency is related to people's existence as individuals, the concept of communion indicates the individual's participation in a larger group. Agency includes self-protection, self-assertion, self-expansion, self-control, and self-direction and emphasizes forming separations. Communion includes group participation, cooperation, attachment, and social connections. It emphasizes the creation of social unions.

We can refer to the Basic Psychological Needs Theory (BPNT) [410, 411] to understand the centrality of agency and communion in human well-being. This theory proposes that three psychological needs promote well-being: autonomy, competence and relatedness [412, 413]. People's need for autonomy indicates their desire to be causal agents and experience volitional behavior [414]. Regarding competence, individuals need to be effective when interacting with the environment [415]. Furthermore, the need for relatedness is manifested through people's inclination to be connected and interact with their peers. It is not difficult to appreciate that autonomy and competence are closely related to the sense of control provided by the concept of agency, while relatedness is directly connected with communion. Agency and communion have been studied in relation to well-being [416] and to depressive styles [417]. Individuals with depressive symptoms reported a reduced sense of control [418]. Both agency and communion were significantly positively associated with positive affect and social adjustment. In fact, in his landmark book on depression [419], clinical psychologist Paul Gilbert concludes

that depression relates to two basic social outcomes, one is power, which is at the core of the agency concept, and the other dimension is a sense of belonging, which is related to communion.

The concepts of agency and communion have been extensively studied in relation to gender stereotypes. In fact, from an empirical approach, these two dimensions were explicitly designed to capture desirable attributes for men and women, respectively [420–422]. Gender gaps in agency and communion are commonly connected with social role theory [423] and to the biosocial extension of social role theory [424]. According to this theoretical framework, these gender gaps emerge when a division of labor by gender is observed, which comes from physical differences between men and women. Gender role expectations are created from the observations of group differences in behavior, which are further used for inferring group differences in traits. They act as normative pressure internalized as aspects of people's identities.

A recent meta-analysis by Hsu et al. [425] estimated that men tended to be more agentic, while women tended to be more communal, consistent with the social role theory [423]. The authors reported that the gender differences between agency and communion decreased with time. These differences decreased with age, but increased with country-level gender segregation in employment. Further, the gender differences in agency and communion also depend on whether or not participants were sampled as couples, and on their sexual orientation [425]. For this reason, it could be expected that gender has a certain impact on the expression of agency and communion in language about happy moments. For estimating the levels of agency and sociality of a post, we develop a semi-supervised classification model that is competitive with other state-of-the-art approaches and use it in our analysis.

To the best of our knowledge, the research on the expression of happiness on social media has not taken gender differences into account in natural language processing. Some work has been done considering the effect of gender on facial expressions of emotion [426, 427] but not in textual data from social media. This gap in NLP research regarding the expression of happiness is probably due to the difficulty of having the gender of users among the most common metadata. Even if some demographics, including gender, were collected in HappyDB, the authors did not explore the gender differences in happy moments [398]. Moreover, we also explore the gender differences in agency and sociality expressed in happy moments. We incorporated two approaches for estimating the gender of the author of a post: a vocabulary-based method searching for specific mentions of gender, and a deep learning method exploiting the writing style of Reddit users.

## 6.3 Extracting Happy Moments from Social Media Data

To address **RQ5 (a)**, we develop a method that can extract the expressions of happy moments from texts written by individuals with or without mentions of a depression diagnosis in social media. Since there is no available happy moments dataset that

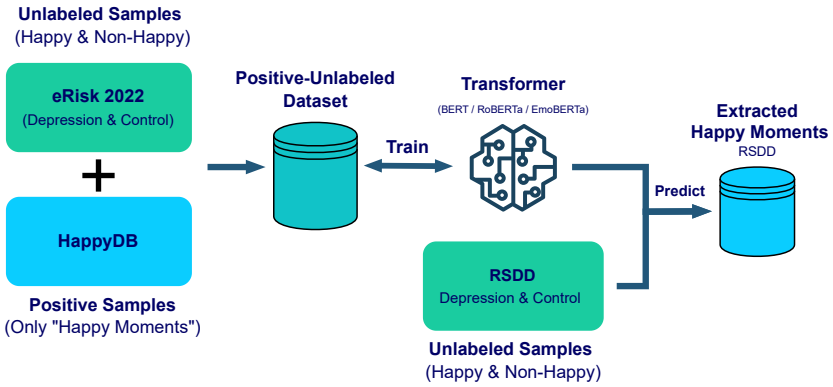



Figure 6.1: Diagram of our method for extracting happy moments from controls and individuals with mentions of depression diagnoses from their social media posts. After fine-tuning a pretrained transformer model under the PU-learning framework, we use it to extract happy moments from RSDD, a distinct large-scale Reddit dataset, to further conduct our analyses.

also includes a negative class, which would allow for direct training of detection models, we formulate the problem in the context of the positive-unlabeled (PU) learning framework. Identification problems are classical PU-learning problems, for which this framework has proven successful.

## Datasets

Our aim is to extract happy moments from RSDD [74], one of the largest social media datasets with Reddit posts from individuals mentioning a depression diagnosis and controls, to be further analyzed for obtaining insights into the way people express their happiness. Since the purpose of RSDD is to be used for the final analyses, we focused on building a PU-dataset containing similar statements to those from RSDD. Consequently, we use a similar dataset with Reddit posts, eRisk 2022 [267], for constructing the PU-dataset. To this end, we combine HappyDB and eRisk to form our positive-unlabeled dataset: we use HappyDB for the positive statements and the training set of eRisk for the unlabeled statements. Figure 6.1 showcases a high-level overview of our method.

**HappyDB** [398] is a dataset with 100,000 happy moments collected from crowd workers' responses to questions about what made them happy in the past day or the past 3 months. It contains a great variety of happy moments about sport (e.g., "I went to the gym this morning and did yoga.", "I have run for 6 miles and the feeling is great."), food (e.g., "I had really good carbonara chicken for dinner.", "I had Hawaiian food for dinner."), nature (e.g., "Going to the mountains and hiking the trail to the top.", "I got wet by a spring rain."), time spent with family (e.g., "We enjoyed our week end with my family", "This morning my nephew and sister in

Table 6.1: Overview of the datasets used for happy moments analysis.

| Dataset | Short Description | Dataset Size | Role in this study |
|---|---|---|---|
| HappyDB [398] | Crowd-sourced happy moments from a diverse range of topics | 100,000 happy moments | Labeled positive class for the PU-dataset |
| eRisk train. [267] | Reddit dataset with users with mentions of a depression diagnosis (the depression group), and without mentions of a depression diagnosis (control group) | 160 users in the depression group and 1105 users in the control group, with an average of 252 posts per user | Unlabeled sentences for the PU-dataset |
| eRisk val. [267] | Reddit dataset with users with mentions of a depression diagnosis (the depression group), and without mentions of a depression diagnosis (control group) | 54 users in the depression group and 371 users in the control group, with an average of 238 posts per user | Manually annotated subset used for validation of happy moment extraction. |
| CL-Aff [401] | Happy moments partially annotated with agency and sociality labels | Training data: 10,560 labeled and 59,846 unalabeled happy moments. Testing data: 17,215 happy moments | Training a semi-supervised model that estimates measures of agency and sociality from happy moments |
| RSDD [74] | Large-scale Reddit dataset with users with mentions of a depression diagnosis (the depression group), and without mentions of a depression diagnosis (control group) | 9,210 users in the depression group and 107,274 users in the control group, with an average of 969 posts per user | Target dataset from which in-the-wild happy moments are extracted |

law came to visit, along with my mother in law.") or friends (e.g., "I went out with friends to have drinks at a bar.", "A friend took me to lunch today."), etc. We use the clean version of HappyDB, in which spelling errors were corrected. Given that HappyDB contains noisy data [397] ranging from very short texts with only two words to long essays with over 1,000 words, the happy moments containing less than 5 or more than 50 words were removed. The duplicate texts were also removed. We used the remaining 90,641 texts from the 100,535 original happy moments in HappyDB.

The **eRisk 2022** [267] and the **RSDD** [74] datasets contain texts from Reddit users who mention that they have a depression diagnosis. The datasets were presented in Seection 2.5.

Because the happy moments from HappyDB are expressed in just one or a few sentences, and the posts from RSDD and eRisk can have up to 40,000 characters (character upper limit from Reddit), we chose to split them into sentences. We pre-process the sentences from RSDD and eRisk by removing URLs, sentences with less than 10 and more than 50 words, and posts that were expressed as questions or in the form of a list of items. We used polyglot Al-Rfou', Perozzi and Skiena [428]

for detecting and removing any non-English sentences.

**PU-learning** [429] has been used in the past in identification problems, in which the aim is to identify examples in a large unlabeled dataset that are similar to a group of positively labeled examples. In real-world applications, positive examples are often more straightforward to acquire than negative ones. Applications such as disease gene identification [430, 431], marketing [432], deceptive opinions detection [433], recommendation systems [434] greatly benefit from the positive-unlabeled framework.

The goal of PU learning [429] is to train a classifier that can distinguish between a positive class and a negative class. However, during the learning phase, only some of the positive examples are labeled, while none of the negative examples are. Formally, a PU dataset is a set of triples $(x, y, s)$, with $x$ being the feature vector, the class $y$ being the class label, and $s$ being a binary variable that indicates whether the tuple $(x, y)$ is labeled or not. In PU learning, it is always the case that, if $s = 1$, then $x$ belongs to the positive class: $P(y = 1|s = 1) = 1$. If $s = 0$, then $x$ can belong to either class. In order to construct a model that learns from positive and unlabeled data, important assumptions are to be made about the labeling mechanism, which refers to the way in which labeled positive samples are selected. One such assumption is the "Selected Completely At Random" (SCAR), which assumes the labeled positive samples are uniformly distributed among all the positive samples [399]. The SCAR assumption enables the reduction of PU learning to a standard binary classification, where all unlabeled samples are treated as negative.

Arguably, statements from HappyDB are qualitatively different from Reddit texts in terms of context and the form in which they were written, and the SCAR assumption might not hold. A more appropriate assumption would be the probabilistic gap assumption $\Delta P(x) = P(y = 1|x) - P(y = 0|x)$: unlabeled samples similar to the negative class are less likely to be labeled [435], and so the labeled positive samples are a biased distribution of total positive samples. This implies that some expressions of happiness found in eRisk are written in a different manner than the statements from HappyDB. Nevertheless, for downstream analysis using the agency and sociality concepts, our models are trained on statements from CL-AFF, a dataset with similar statements to those in HappyDB. Consequently, to avoid further downstream issues regarding the domain gap, we are interested in statements in eRisk, which are written in the same manner as those in HappyDB, essentially imposing the SCAR assumption. As a consequence, under the SCAR assumption, we can train a binary classification model considering the unlabeled examples pertaining to the negative class [399].

Under the PU-learning framework, we compare three different variants of PU-learning for the identification of happy moments.

**Binary Classification** In this training scenario, we directly train a model considering the statements in eRisk to be the negative class, and statements from HappyDB to be the positive class. We do not do any additional re-weighting on the class probabilities.

**Elkanoto** Following the SCAR assumption, the positive class probability from the binary classifier trained above can be adjusted [399] by taking into account the class prior $P(y = 1|x) = \frac{P(s=1|x)}{c}$, where $c = P(s = 1|y = 1)$ is the label frequency.

**Weighted Elkanoto** The probability for an example to be positive can be further adjusted [399] by incorporating the label probabilities: $P(y = 1|s = 0, x) = \frac{1-c}{c} \cdot \frac{P(s=1|x)}{1-P(s=1|x)}$, where $P(s = 1|x)$ is the probability for a sample to be positive, given by the trained binary classifier.

For our experiments, we implemented the PU-learning framework using three pretrained transformer models: BERT [335] and RoBERTa [327], for their widespread use and general expressive power, and EmoRoBERTa [436], a RoBERTa model fine-tuned on the GoEmotions [437] dataset. While happy moments have been shown to not necessarily be correlated with positive emotions [395], we chose to also include EmoRoBERTa for detecting happy moments, since, nevertheless, the task is dealing with affective statements.

For each pretrained model, we froze the backbone and fine-tuned only the final layer for an epoch, and subsequently unfroze the rest of the layers for 3 more epochs. We trained all models with an AdamW optimizer [294], with a learning rate of 0.000001, following a linear schedule with decay. The training set is severely imbalanced since eRisk contains one order of magnitude more sentences than HappyDB. We downsample the number of sentences from eRisk by factors of 0.2 and 0.5, respectively, corresponding to class imbalances of 1:2 and 1:4.

**Emotional Context Filtering** Given that our methodology for extracting happy moments requires the texts to be short, not as long as a full Reddit post (which has a 40,000 characters limit), the originating context of the candidate happy moment statement must be taken into account when formulating a decision, in order to avoid false positives. Since a manual analysis of all the extracted happy moments and their context was unfeasible, we opted for an automatic method, extracted fine-grained emotion labels for the happy moments and the posts they originate from, and filtered out the happy moments in a negative context. The emotions were extracted using EmoRoBERTa [436], a RoBERTa model trained on the GoEmotion dataset [437] to provide fine-grained emotion labels for text. GoEmotion is a large human-annotated dataset containing data collected from Reddit and annotated with 28 emotion labels, including neutral. Each happy moment and its context were labeled with the predominant emotion predicted by EmoRoBERTa.

## Evaluation

For evaluating the proposed method for extracting happy moments, a subset of 1200 sentences from the eRisk validation dataset was annotated by 3 expert annotators. The annotators were instructed to mark whether or not a given text represents a happy moment. In line with the research performed by Asai et al. [398], we consider a happy moment to be the natural language expression of an experience that brings feelings of joy, contentment, and satisfaction to an individual. The annotators received examples of statements from HappyDB [398] to better understand how

happy moments are expressed. The annotators had a substantial inter-rater agreement (Fleiss' $\kappa = 0.65$). The final label for each manually annotated text was decided through majority voting, totaling 607 positive samples annotated as representing a happy moment and 593 negative samples.

Table 6.2: Results for happy moments extraction using Positive-Unlabeled Learning, with different methods to adjust prior probabilities. We downsampled the number of sentences from eRisk to balance out the classification labels. With bold and italic we mark the best and second best results per column, respectively.

| Method | frac. of eRisk | Sentence Encoder | Accuracy | Precision | Recall | $F_1$ |
|---|---|---|---|---|---|---|
| **Binary Classification** | **0.2** | BERT | 0.7549 | 0.7981 | *0.6909* | *0.7407* |
| | | RoBERTa | 0.7082 | 0.8180 | 0.5453 | 0.6544 |
| | | EmoRoBERTa | 0.7549 | 0.8038 | 0.6828 | 0.7384 |
| | **0.5** | BERT | 0.7107 | 0.8132 | 0.5566 | 0.6609 |
| | | RoBERTa | 0.7008 | **0.8338** | 0.5113 | 0.6339 |
| | | EmoRoBERTa | 0.7344 | 0.8050 | 0.6278 | 0.7055 |
| **Elkanoto** | **0.2** | BERT | 0.7549 | 0.7981 | *0.6909* | *0.7407* |
| | | RoBERTa | 0.7082 | 0.8180 | 0.5453 | 0.6544 |
| | | EmoRoBERTa | *0.7557* | 0.8042 | 0.6845 | 0.7395 |
| | **0.5** | BERT | 0.7115 | 0.8122 | 0.5599 | 0.6628 |
| | | RoBERTa | 0.7000 | *0.8316* | 0.5113 | 0.6333 |
| | | EmoRoBERTa | 0.7352 | 0.8041 | 0.6310 | 0.7072 |
| **Weighted Elkanoto** | **0.2** | BERT | **0.7631** | 0.7975 | **0.7136** | **0.7532** |
| | | RoBERTa | 0.7115 | 0.8152 | 0.5566 | 0.6615 |
| | | EmoRoBERTa | 0.7549 | 0.8038 | 0.6828 | 0.7384 |
| | **0.5** | BERT | 0.7122 | 0.8027 | 0.5728 | 0.6686 |
| | | RoBERTa | 0.7041 | 0.8303 | 0.5227 | 0.6415 |
| | | EmoRoBERTa | 0.7336 | 0.8046 | 0.6262 | 0.7043 |

## Results

For answering **RQ5 (a)**, we show results for happy moment extraction in Table 6.2. Across the different probability weighting schemes, the performance change is marginal. Increasing the fraction of eRisk proved detrimental to the performance, due to increased imbalance between labeled and unlabeled data. EmoRoBERTa achieved reasonable performance, especially compared to a plain RoBERTa model, indicating that affective features are useful for this task. However, for the final analysis, we used the BERT model trained using Weighted Elkanoto scaling and a fraction of 0.2 of eRisk, achieving the best overall scores. Using the aforementioned model, we extract happy moments of both the depression and control groups from the entire RSDD dataset (train, validation, and test splits). We obtained 498,551 moments, of which 128,857 were from the depression group and 369,694 were from the control users.

After extracting the data, we filtered out the extracted happy moments originating from contexts with negative emotions (i.e., fear, sadness, anger, annoyance, disappointment, disapproval, disgust, embarrassment, grief, nervousness, and remorse)

Table 6.3: Examples of extracted happy moments and their corresponding context. We filter out happy moments originating from negative emotional context. While most of the happy moments have neutral or positive emotions, the originating context provides additional disambiguation of the statement meaning.

| **Extracted Happy Moment** | **Context** | **Emotion Happy Moment** | **Emotion Context** | **Alignment** |
|---|---|---|---|---|
| Yesterday for lunch I ordered a French Dip with fries at a pretty fancy restaurant. | They were the worst fries I'd ever had. | 😌 joy | 😖 disgust | ❌ |
| The Rotary Exchange Program chose me to be an exchange student in Japan. | My parents could not afford to let me go. They gave the placement to another person who could afford it. | 👍 approval | 😢 sadness | ❌ |
| My aunt gave me a six-pack of VHS cassettes. | I recorded a "video" with my then boyfriend, which his aunt and uncle discovered. I was fifteen. The worst present ever! | 😐 neutral | 😖 disgust | ❌ |
| I arrived at my office early to get some work done. | I heard my coworker, who works in my department, at her cubicle and had a question. [...] I stroll over and find her editing sexual photographs of herself. | 👍 approval | 😳 embarrassment | ❌ |
| I applied for what appeared to be a dream job. | The salary was great. [...] However, the position would be a career killer. When I approached my manager, he laughed at how he had deceived me. I quit. | 👍 approval | 😒 annoyance | ❌ |
| My friend wrote an extremely detailed and long entry in my yearbook. | It was amazing [...] The whole thing made me tremendously happy. | 😐 neutral | 😌 joy | ✅ |
| My wife recently got me a new watch for my birthday! | [...] The chronograph and date window are nice, and it looks really good on my wrist! [...] | 😌 joy | 👏 admiration | ✅ |
| I pushed myself to go for a walk and enjoy the weather and didn't wear any makeup! | So it was a successful day. Really turned around from drinking a bottle of wine impulsively. | 😌 joy | 🙏 gratitude | ✅ |

predicted by the EmoRoBERTa model. In Table 6.3, we show a selection of happy moments, the post they originate from, and the predicted emotion labels. For the first five rows, there are misalignments in the emotion of the happy moment and the broader context. In the first columns, there are sentences that, in appearance, seemed to be happy moments, but after analyzing the context they originate from, they were deemed false positives. In the case of the last three rows, the emotions from the happy moment and the context were consistent and deemed true positives. Using emotion filtering, 16.09% of happy moments from the depression group and 11.83% from the control group were removed. The number of samples was reduced to 108,125 happy moments from the depression group and 325,948 happy moments from the control group.

## 6.4 Inferring Agency and Sociality

To address **RQ5 (b)**, we develop a semi-supervised method to predict agency and sociality using labeled and unlabeled data. For this purpose, we use the CL-Aff HappyDB dataset Jaidka et al. [395, 401] as a training dataset.

### Dataset

The **CL-Aff HappyDB** [395, 401] dataset builds upon HappyDB [398]. The dataset contains 10,560 happy moments from the original HappyDB manually annotated for agency and sociality. The annotations were made by crowd-workers who received instructions following the Agency-Communion framework developed by Paulhus and Trapnell [400]. The annotators were asked if the author is in control of a specific happy moment (agency) and if the moment involves other people than the author (sociality). They were provided with examples of happy moments in which an author has agency (e.g., "I ran on the treadmill for 20 minutes straight when I could barely do 5 minutes 3 months ago.") or not (e.g., "My youngest daughter got accepted to many prestigious universities and accepted an offer to attend college in San Diego.") and happy moments that are social (e.g., "Going out to a special birthday lunch for my great-grandmother in law's birthday.") or not (e.g., "The bus came on time, so I reached work early.") [401]. Besides the labeled data from training and testing, an additional unlabeled training set with 59,846 texts from HappyDB is provided. The CL-Aff HappyDB test set dataset is comprised of 17,215 new happy moments annotated in a similar manner.

### Methodology

We propose fine-tuning a RoBERTa transformer in a multitask setting to predict agency and sociality from CL-AFF data. That is, we attempt to predict both agency and sociality of a happy moment from a single forward pass. To achieve this, we leveraged the manually labeled data alongside the unlabeled statements,

and implemented the Unsupervised Data Augmentation [438] approach for semi-supervised learning.

**Unsupervised Data Augmentation (UDA)** [438] is a general state-of-the-art method for semi-supervised learning, which is directly applicable to textual data. The method uses consistency regularization on the unlabeled examples. Given a classification network $f_\theta(\cdot)$ and given labeled input $x_l$, the predicted class estimated by $\tilde{y} = f_\theta(x_l)$ is compared to real label $y$, as in classical supervised classification, using cross-entropy loss. Consistency regularization uses the unlabeled examples $x_u$ through random augmentation to compute a divergence loss (in our case cross-entropy) between the original and corrupted version $\tilde{x}_u$ of the example: $CE(f_\theta(\tilde{x}_u), f_{\tilde{\theta}}(x_u))$, where $f_{\tilde{\theta}}$ denotes a fixed copy of $f_\theta$, through which gradients are not propagated. In other words, the prediction for the unlabeled example should remain the same, irrespective of augmentation. The loss for a classification output $i$ is then:

$$\mathcal{L}_\theta^{(i)} = CE(f_\theta(x_l), \tilde{y}) + \lambda CE(f_\theta(\tilde{x}_u), f_{\tilde{\theta}}(x_u))$$

Where $\lambda$ is a weighting parameter that controls regularization strength. For a multitask setting, the same loss is applied for each classification output, giving $\mathcal{L}_\theta = \frac{1}{2}(\mathcal{L}_\theta^{(agency)} + \mathcal{L}_\theta^{(sociality)})$.

In practice, for augmentations, we follow the authors' implementation [438] and we used back-translations from English to French, using a pair of pretrained multilingual models[3][4] trained on OPUS data [439]. Furthermore, we also include random insertion of contextual word embeddings using a pretrained RoBERTa masked language model.

For this method, we trained the model for 10 epochs, using a batch size of 16, and a learning rate of 0.00001, using a linear decay scheduler, with the AdamW [294] optimizer. We froze the RoBERTa backbone for the first epoch and only fine-tuned the classification heads. Afterwards, the whole network is trained.

We compare our approach with state-of-the-art methods from the participating teams in the CL-Aff Shared Task, such as LSTM using ELMo embeddings [440], Averaged SGD Weight-Dropped LSTM (AWD-LSTM) [441], Word Pair Convolutional Models (WoPCoM) [442], Recurrent Convolutional Neural Networks (RCNN) [443] and Convolutional Neural Networks (CNN) with GloVe features [444].

### Results

For answering **RQ5 (b)**, in Table 6.4, we showcase our results on CL-AFF data for the estimation of agency and sociality from happy moments. Our methods are on

[3] https://huggingface.co/Helsinki-NLP/opus-mt-en-fr.
[4] https://huggingface.co/Helsinki-NLP/opus-mt-fr-en.

Table 6.4: Results for semi-supervised classification of agency and sociality from CL-Aff dataset.

| | Agency | | | | Sociality | | | |
|---|---|---|---|---|---|---|---|---|
| **Method** | **Accuracy** | **$F_1$** | **Precision** | **Recall** | **Accuracy** | **$F_1$** | **Precision** | **Recall** |
| LSTM [440] | 0.85 | 0.90 | - | - | 0.92 | 0.93 | - | - |
| AWD-LSTM [441] | 0.84 | 0.89 | - | - | 0.92 | 0.93 | - | - |
| WoPCoM [442] | 0.85 | 0.89 | - | - | 0.91 | 0.92 | - | - |
| RCNN [443] | 0.83 | 0.89 | - | - | 0.91 | 0.92 | - | - |
| CNN [444] | 0.83 | 0.88 | - | - | 0.89 | 0.90 | - | - |
| RoBERTa Supervised | 0.8539 | 0.8962 | 0.8995 | 0.8929 | 0.9320 | 0.9400 | 0.9439 | **0.9362** |
| RoBERTa UDA | 0.8611 | **0.9019** | 0.8995 | **0.9043** | **0.9339** | **0.9415** | 0.9487 | 0.9344 |

Table 6.5: Selected examples of happy moments with their corresponding predictions for agency and sociality.

| **(+) Agency & (+) Sociality** | **(+) Agency & (–) Sociality** | **(–) Agency & (+) Sociality** | **(–) Agency & (–) Sociality** |
|---|---|---|---|
| I helped a friend to repair his relationship with his girlfriend. | I recently upgraded to a new smartphone, which is great. | I was thrilled when my youngest pushed around his car. | It made my day when my doorbell rang with my package! |
| I kissed the girl who is my biggest crush. | I recently bought two 35mm film cameras, which I absolutely love. | My oldest son tied his shoes for the first time! | I was invited to work in the United States for a month. |
| I went to mother's birthday party yesterday evening and it was nice. | I won first place in the teen bodybuilding competition. | When the Nintendo Wii was out, my father bought me one for Christmas. | After a rocky start, my cat "treated" my moderate allergy through exposure! |

par with the other state-of-the-art works and surpass them by a small margin in some cases. The methods proposed for the CL-Aff Shared Task do not report precision and recall values, but we included the measures for our models for completeness. We use RoBERTa with Unsupervised Data Augmentation for further analyses since it achieves 0.90 $F_1$ for agency and 0.94 $F_1$ for sociality, the highest overall scores.

In Table 6.5, we showcase various happy moments extracted from RSDD labeled for agency and sociality using the previously mentioned model. Three examples are displayed for each combination of agency and sociality labels (e.g., happy moments that convey both agency and sociality, examples that convey agency and not sociality, and other variants).

## 6.5 Topics of Happy Moments

To address **RQ5 (c)**, we further detail the mechanism for topic extraction from the RSDD dataset (the description of the RSDD dataset can be found in Section 6.3), using BERTopic [445]. For inferring gender, we used a hybrid approach based on self-reported gender and automatically inferred gender [403].

## Methodology

We opted to use an online topic modeling approach, which incrementally learns from mini-batches of instances, since RSDD contains several million texts, and processing all of them at once is computationally unfeasible.

**BERTopic** generates topics in three steps: (i) document embeddings are generated using pre-trained sentence transformer models (in our case, RoBERTa), (ii) dimensionality reduction is performed using incremental Principal Component Analysis (PCA), and documents are clustered using Mini-Batch K-Means, and (iii) topic representations for each cluster are extracted using class-based Term Frequency - Inverse Document Frequency (TF-IDF). As opposed to other topic modeling approaches [446, 447], which use keyword bag-of-words representations that ignore the semantic relations between words, BERTopic uses contextual document embeddings and provides high-level semantic topics. We extracted 50 topics from RSDD. After computing the similarity between these topics, and manually inspecting them, we merged the most similar topics, and reduced the total number of topics to 30. We chose suitable names for each topic upon manual examination.

We use UMAP McInnes et al. [448] to visualize topic clusters to perform supervised dimensionality reduction. We train a UMAP model with 20 neighbors on the document embeddings for all happy moments.

**Gender inference** To analyze gender differences between users, we automatically inferred the gender of the users from RSDD and used it as an independent variable in our analysis, together with the depression label. Following prior research [449–451], we infer binary gender labels, due to the lack of more diverse data or models being available for this purpose. For inferring the gender labels, we use a hybrid approach adapted from Jagfeld et al. [452]:

**Self-reported gender** We extracted the gender using regular expressions of users from their mentions (e.g., "I am a woman", "I'm a man") or from the standard bracketed format used for self-reports on Reddit (e.g., "I [24f]", "my [25/m]"). Using this method, we extracted the gender of 11% of users from our data. We also use this subset as an additional testing set for the computational approach described below.

**Automatically inferred gender** For the users without self-reported gender, we used the state-of-the-art CNN and attention-based model $\text{HAM}_{CNN-attn}$ from Tigunova et al. [403] trained on the Reddit data labeled for gender from the RedDust dataset [449]. The model uses a maximum of 100 posts from each user to infer the gender. We trained the model using the publicly available implementation provided by the authors[5], with the same hyper-parameters. The data from the RedDust dataset was split using the same 9:1 proportion for train and test sets, as in the original implementation. We evaluated the performance of the $\text{HAM}_{CNN-attn}$ model on the test set and also against the user gender labels retrieved through regular expressions (omitting the posts with explicit mentions of gender). We trained the state-of-the-art

[5] https://github.com/Anna146/HiddenAttributeModels.

CNN and attention-based model $\text{HAM}_{CNN-attn}$ and obtained 0.82 Accuracy and 0.82 F1-score, as opposed to 0.85 Accuracy reported by the authors [449]. This discrepancy in performance might be due to a different random splitting of the data using the same 9:1 proportion for train and test sets. We used the trained model to infer the gender of the users from RSDD. Given that our users have more than 100 posts, we split the user's timeline into multiple sequences of 100 posts and predicted the gender for each sequence. The final gender label of the user is computed through majority voting. We also evaluated the performance of the $\text{HAM}_{CNN-attn}$ model against the user gender labels retrieved through regular expressions. We obtained a performance of 0.86 Accuracy and 0.86 F1-score. Furthermore, the model achieved a balanced performance for each class: 0.88 F1 for females and 0.84 F1 for males. We consider these results to be of acceptable performance, which allows us to conduct our analysis further.

**Statistical Tests** Mann-Whitney U Test was used to measure the differences in happy moments ratios between users in the depression and the control groups. Chi-square test is used for conducting statistical analysis on the topics of happy moments.

## Results

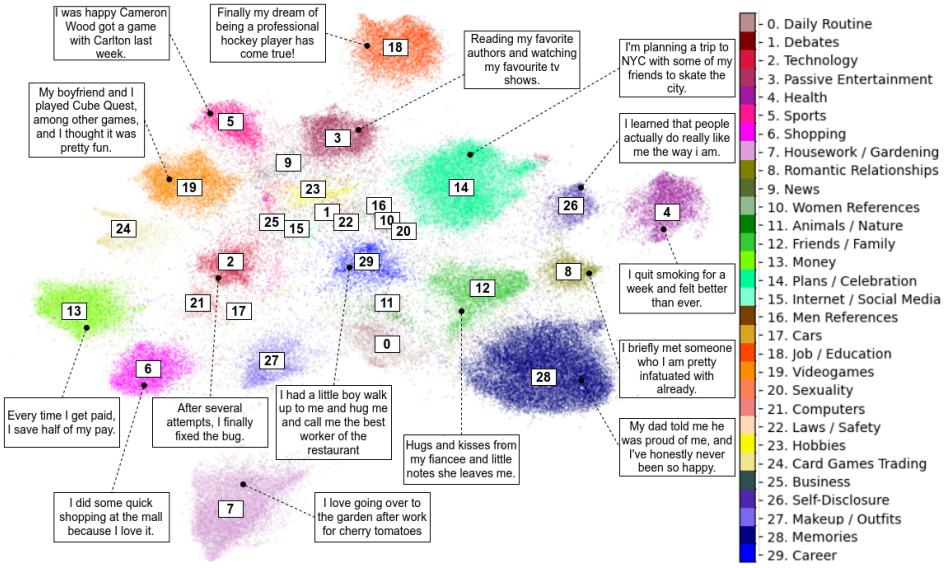



Figure 6.2: Topic clusters of happy moments represented using supervised dimensionality reduction with UMAP. Examples are paraphrased to protect the users' privacy.

Figure 6.2 showcases the topic clusters for all the happy moments extracted from RSDD, alongside selected examples that were paraphrased. In the center of the figure, there are topics that have very few happy moments, such as Business,

Debates, Sexuality, and others. The most predominant topics in the happy moments are Memories, Plans/Celebration, and Housework/Gardening.

By inferring the gender of the users, we obtained that men have more happy moments in the control group than women: 76.61% out of the 325,948 happy moments. For the depression group, there are more happy moments written by women: 60.36% of 108,125 happy moments. These results are consistent with the demographic skew of Reddit towards male users [453] and with the current research on depression, which finds that this mental disorder is more prevalent in women than in men [454]. The prevalence of happy moments (the ratio of happy moments to the total number of posts for each user) was 0.0074 for females in the depression group and 0.0114 for females in the control group. For male users, the prevalence was 0.0053 for the depression group and 0.0072 for the control group. As the ratio of happy moments does not follow a normal distribution, Mann-Whitney U Test was used to measure the differences between users in the depression and the control groups. For both genders, the control group registers a significantly greater ratio of happy moments than the depression group (Mann-Whitney U=302964771.50; $p$<.001). There is also a significant effect of gender: women register a greater ratio of happy moments than men (Mann-Whitney U=466330044.50; $p$<.001)

For further analyses, we balanced the happy moments of individuals in the depression and control groups by gender. We randomly selected 42,863 posts (the minimum number of happy moments from male users in the depression group) from each group: females in the depression group, males in the depression group, females in the control group, and male users in the control group.

To answer **RQ5 (c)**, in Table 6.6, we show the differences in gender in the topics of happy moments with more than 2,000 instances. The differences in topics between male and female users are statistically significant (Pearson $\chi^2$=10049.701, df=29, $p$<.001), the topics with a standardized residual beyond the range of ±2.58 are considered to have a major contribution to the Chi-square significance. Male users express more happy moments related to daily routine, job/education, money, passive entertainment, sports, technology and videogames, while female authors express more happy moments related to friends/family, health, housework/gardening, makeup/outfits, plans/celebrations, romantic relationships, memories and shopping. The gender variable does not introduce significant differences when the topic of the happy moment is about career or self-disclosure.

Due to these gender differences in the topics of happy moments, we compare the people in the depression group with the control group, separating the users by gender. In Table 6.7, we showcase the results for the topic differences in happy moments between depression and control groups for both male and female users. The topics of male authors in the depression group differ significantly (Pearson $\chi^2$= 3240.221, df=29, $p$<.001) from the male control group. The topics with a standardized residual beyond the ±2.58 contribute to the Chi-square significance. As shown in Table 6.7, the happy moments of males in the depression group focus more on career,

Table 6.6: Gender per topic for happy moments. Bold denotes that there are significantly ($p < 0.01$) more happy moments in a given topic in one of the groups.

| **Topic** | | **Male** | **Female** |
|---|---|---|---|
| Career | Count | 2429 | 2416 |
| | Adjusted Residual | 0.2 | -0.2 |
| Daily Routine | Count | 1928 | 1481 |
| | Adjusted Residual | **7.7** | -7.7 |
| Friends / Family | Count | 4647 | 5256 |
| | Adjusted Residual | -6.3 | **6.3** |
| Health | Count | 3545 | 5640 |
| | Adjusted Residual | -22.5 | **22.5** |
| Housework / Gardening | Count | 6648 | 10891 |
| | Adjusted Residual | -33.8 | **33.8** |
| Job / Education | Count | 5690 | 4882 |
| | Adjusted Residual | **8.1** | -8.1 |
| Makeup / Outfits | Count | 1611 | 3744 |
| | Adjusted Residual | -29.6 | **29.6** |
| Money | Count | 5330 | 3747 |
| | Adjusted Residual | **17.1** | -17.1 |
| Passive Entertainment | Count | 4478 | 2658 |
| | Adjusted Residual | **22.0** | -22.0 |
| Plans / Celebration | Count | 9875 | 10333 |
| | Adjusted Residual | -3.4 | **3.4** |
| Romantic Relationships | Count | 1420 | 2882 |
| | Adjusted Residual | -22.6 | **22.6** |
| Self-Disclosure | Count | 2017 | 1933 |
| | Adjusted Residual | 1.4 | -1.4 |
| Memories | Count | 15142 | 18849 |
| | Adjusted Residual | -22.5 | **22.5** |
| Shopping | Count | 3049 | 3411 |
| | Adjusted Residual | -4.6 | **4.6** |
| Sports | Count | 3597 | 817 |
| | Adjusted Residual | **42.4** | -42.4 |
| Technology | Count | 2661 | 992 |
| | Adjusted Residual | **27.9** | -27.9 |
| Videogames | Count | 5357 | 1425 |
| | Adjusted Residual | **48.7** | -48.7 |

FRIENDS/FAMILY, HEALTH, JOB/EDUCATION, ROMANTIC RELATIONSHIPS, MEMORIES and SELF-DISCLOSURE. The male users from the control group express more happy moments focusing on MONEY, PLANS/CELEBRATIONS, SHOPPING, SPORTS, TECHNOLOGY and VIDEOGAMES. There are no significant differences related to the DAILY ROUTINE, HOUSEWORK/GARDENING, MAKEUP/OUTFITS, and PASSIVE ENTERTAINMENT in the happy moments of and male users in the depression and control groups.

Regarding the female gender, the topics of happy moments from female users in the depression group differ significantly (Pearson $\chi^2$=1690.497, df=29, $p$<.001) from the female authors from the control group. The happy moments expressed by

females in the depression group focus more on CAREER, FRIENDS/FAMILY, HEALTH, JOB/EDUCATION, ROMANTIC RELATIONSHIPS, MEMORIES and SELF-DISCLOSURE. They express fewer happy moments in comparison with the control group, focusing

Table 6.7: Happy moments of depressed and control users by gender. Bold denotes that there are significantly ($p < 0.01$) more happy moments in a given topic in one of depression or control groups, for each gender.

| | | Male | | Female | |
|---|---|---|---|---|---|
| **Topic** | | **Control** | **Depression** | **Control** | **Depression** |
| Career | Count | 1117 | 1312 | 1044 | 1372 |
| | Adjusted Residual | -4.0 | **4.0** | -6.8 | **6.8** |
| Daily Routine | Count | 981 | 947 | 774 | 707 |
| | Adjusted Residual | 0.8 | -0.8 | 1.8 | -1.8 |
| Friends / Family | Count | 1806 | 2841 | 2258 | 2998 |
| | Adjusted Residual | -15.6 | **15.6** | -10.5 | **10.5** |
| Health | Count | 1191 | 2354 | 2389 | 3251 |
| | Adjusted Residual | -19.9 | **19.9** | -11.9 | **11.9** |
| Housework / Gardening | Count | 3391 | 3257 | 5741 | 5150 |
| | Adjusted Residual | 1.7 | -1.7 | **6.1** | -6.1 |
| Job / Education | Count | 2208 | 3482 | 2117 | 2765 |
| | Adjusted Residual | -17.5 | **17.5** | -9.6 | **9.6** |
| Makeup / Outfits | Count | 831 | 780 | 2232 | 1512 |
| | Adjusted Residual | 1.3 | -1.3 | **12.0** | -12.0 |
| Money | Count | 3014 | 2316 | 2016 | 1731 |
| | Adjusted Residual | **9.9** | -9.9 | **4.8** | -4.8 |
| Passive Entertainment | Count | 2299 | 2179 | 1489 | 1169 |
| | Adjusted Residual | 1.8 | -1.8 | **6.3** | -6.3 |
| Plans / Celebration | Count | 5169 | 4706 | 5542 | 4791 |
| | Adjusted Residual | **5.0** | -5.0 | **7.9** | -7.9 |
| Romantic Relationships | Count | 491 | 929 | 1206 | 1676 |
| | Adjusted Residual | -11.7 | **11.7** | -8.9 | **8.9** |
| Self-Disclosure | Count | 629 | 1388 | 745 | 1188 |
| | Adjusted Residual | -17.1 | **17.1** | -10.2 | **10.2** |
| Memories | Count | 6784 | 8358 | 8762 | 10087 |
| | Adjusted Residual | -14.1 | **14.1** | -10.9 | **10.9** |
| Shopping | Count | 1813 | 1236 | 2084 | 1327 |
| | Adjusted Residual | **10.6** | -10.6 | **13.2** | -13.2 |
| Sports | Count | 2507 | 1090 | 468 | 349 |
| | Adjusted Residual | **24.1** | -24.1 | **4.2** | -4.2 |
| Technology | Count | 1625 | 1036 | 519 | 473 |
| | Adjusted Residual | **11.6** | -11.6 | 1.5 | -1.5 |
| Videogames | Count | 3430 | 1927 | 878 | 547 |
| | Adjusted Residual | **21.2** | -21.2 | **8.8** | -8.8 |

ON HOUSEWORK/GARDENING, MAKEUP/OUTFITS, MONEY, PASSIVE ENTERTAINMENT, PLANS/CELEBRATIONS, SHOPPING, SPORTS, and VIDEOGAMES. There are no significant differences in the happy moments related to DAILY ROUTINE and TECHNOLOGY.

## 6.6 Sociality and Agency in the Expression of Happy Moments

To address **RQ5 (d)**, we use the binary labels for agency and sociality to examine the variations between the depression and control groups, based on the gender labels inferred.

### Methodology

We use the binary labels for agency and sociality from the RoBERTa model with Unsupervised Data Augmentation presented in Section 6.4. The methodology for inferring gender labels was presented in Section 6.5. Chi-square test is used for conducting statistical analysis, given the categorical data.

### Results

To answer **RQ5 (d)**, means with confidence intervals for agency and sociality depending on gender and depression status are presented in Figure 6.3. In Table 6.8, we present the results of the Chi-square test, showing statistically significant differences between the control and the depression groups in sociality for men (Pearson $\chi^2$=232,114, df=1, $p$<.001) and women (Pearson $\chi^2$=380,627, df=1, $p$<.001). However, in the expression of agency, the Chi-square is significant among women (Pearson $\chi^2$=111,575, df=1, $p$<.001) and not among men (Pearson $\chi^2$=0,658, df=1, $p$=.42). The latter result might be because agency, an important feature of the male gender stereotype, is not affected by depression. Regarding sociality, both males and females from the depression group have more happy moments expressing sociality than the ones in the control group.

Table 6.8: Agency and Sociality differences in users from the depression and control group using Chi-square test. Bold denotes that there are significant differences (p < 0.01).

| | | **Female** | | **Male** | |
|---|---|---|---|---|---|
| | | **Control** | **Depressed** | **Control** | **Depressed** |
| **Agency** | **(+)** | 32,774 (**10.6**) | 31,433 (-10.6) | 33,744 (-0.8) | 33,841 (0.8) |
| | **(–)** | 10,089 (-10.6) | 11,430 (**10.6**) | 9,119 (0.8) | 9,022 (-0.8) |
| **Sociality** | **(+)** | 19,809 (-19.5) | 22,665 (**19.5**) | 16,923 (-15.2) | 19,125 (**15.2**) |
| | **(–)** | 23,054 (**19.5**) | 20,198 (-19,5) | 25,940 (**15.2**) | 23,732 (-15,0) |

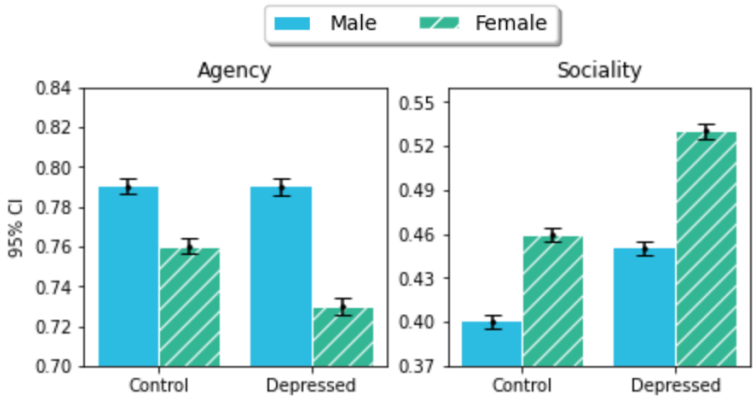



Figure 6.3: Agency and Sociality differences in users from the depression and control group by gender with confidence intervals.

## 6.7 Discussion

### RQ5 (a) and RQ5 (b): Extraction of Happy Moments and Estimation of Agency and Sociality

Using the positive-unlabeled framework, we trained a model that can detect happy moments from texts of individuals with mentions of a depression diagnosis and control users (without any mention). Even if the best-performing model, the BERT model trained using Weighted Elkanoto scaling, performs well in detecting happy moments (0.75 F1 score) on the manually annotated data, an essential step in our post-processing was emotion filtering. Using this approach, false positives were filtered out from further analyses, consequently removing considerable noise from the data (more than 64,000 texts). The first two examples of extracted happy moments (Table 6.3) follow a similar structure to the happy moments from HappyDB [398], and we cannot informatively remove them without looking at the broader context. The model is correctly extracting what seems to be a happy moment, without having access to the whole context. In Table 6.3, we also show a sentence that was incorrectly predicted as being a happy moment ("I applied for what appeared to be a dream job."), in which it can be understood that the experience was not good by the usage of "appeared". Emotion filtering is useful in two situations: a) when a happy moment seems to be happy, but the context is negative, and b) when we have an incorrect prediction from the start.

In line with the works of Moreno-Ortiz et al. [397] and Jaidka et al. [395], our results convey that happy moments are qualitatively different from simple positive sentiment statements, and, consequently, cannot be extracted by solely using emotion detection models. Happy moments can express negative emotions such as sadness (e.g., "My family bought me a puppy after my grandpa asked whom I was playing with, while I was playing alone in the yard.", "I reconnected with a former high

school friend I had lost touch with, after we both posted about a mutual friend who had died."). Furthermore, not every text that expresses a positive emotion is a happy moment (e.g., "Guinea pigs are happiest when they live with at least one other buddy.", "I like what you have done with the floors and the fireplace!").

With regard to agency and sociality, the semi-supervised multi-task method proposed in this work was comparable with the state-of-the-art methods. We achieve high performance in terms of F1 score (0.90 for agency and 0.94 for sociality). The model can correctly predict the sentences without agency when the sentences are also written in the first person (e.g., "I was thrilled when...", "I was invited..."), even if the majority of sentences indicating agency are using first-person pronouns (Table 6.5). Regarding sociality, the model correctly predicts the concept if the happy moment involves other people besides the author (e.g., "my mother", "a friend", "my oldest son"), and is consistent with previous works by not labeling examples with animals with sociality [395]. The models developed in this work for agency and sociality have a good performance, with a higher performance being hard to obtain due to mislabeling of the original data (approx. 10% of label noise for agency and 5% for sociality) [401].

### RQ5 (c): Differences in Topics

The analysis of the topics of happy moments shows the importance of introducing the gender variable. There are differences in topics between male and female users, and, according to our results, these differences also interact with the effect of being or not depressed (**RQ5 (c)**).

After controlling for the gender variable, we observe a common pattern for males and females in the depression group: their happy moments rely significantly more than the control group on topics related to social relationships in personal life (FRIENDS/FAMILY and ROMANTIC RELATIONSHIPS), professional development (CAREER), MEMORIES and personal communication (SELF-DISCLOSURE) when they express happiness. It may be evidence of how social isolation is associated with depressive symptoms [455]. This result also demonstrates the significance of the research showing that people with depression are less likely to spend time with others, even if spending time with close people was associated with an improvement in positive affect [456].

However, in the topics of the happy moments, there are also gender differences that are related to gender stereotypes. For males, we find HOUSEWORK/GARDENING and MAKEUP/OUTFITS among the topics in which depression has no effect. For female users, we find TECHNOLOGY among the topics on which depression has no effect. These results are in line with the idea that experiences of depression are associated with different facets of sex role stereotypes [457, 458].

### RQ5 (d): Differences in Agency and Sociality

The importance of taking gender into account is also evident when we analyze the differences in agency and sociality in the expression of happy moments between users in the depression and the control group (**RQ5 (d)**): in the influence that depression has on the expression of happiness, there are gender commonalities but also important gender differences.

Both male and female users in the depression group find moments of happiness in their social relationships: the sociality dimension is significantly higher than in the control group for both genders. In this sense, our results are coherent with one of the most well-known studies about life satisfaction, the Harvard Study of Adult Development[6], which shows that the social component is an important part of people's happiness. This study concluded that spending more time with others was linked with greater happiness Waldinger and Schulz [459]. Other works have also emphasized the importance of close personal relationships in increasing happiness [460–462]. One possible interpretation of this result is that people with depression are more sensitive to the importance of these social relationships since they are experiencing a moment of vulnerability [463]. From a therapeutic intervention standpoint, this result emphasizes the importance of intervening in the social environment of people suffering from depression.

In the dimension of agency, there are relevant gender differences that have to be taken into account. The cultural assumptions about men having more agency and women being more sociable [425] are reflected in how male and female authors express their happy moments. An important result of our research is that when individuals have depression, their feelings of agency in their moments of happiness are affected differently, depending on their gender. The happy moments of female users in the depression group express less agency than of the female users from the control group. However, for male users, there are not any differences depending on the depression status. Our interpretation is that for women, the absence of control over their lives is a much more problematic issue than for men, and is therefore affected by depression. In summary, our results suggest that some gender stereotypes interplay with the feelings of vulnerability that characterize depression, at least among the female group.

From a therapeutic perspective, these results on agency and sociality expressed in happy moments from individuals with mentions of depression diagnoses could have implications, mainly in the frame of positive psychology (for a review, see the work of McTiernan, Gullon-Scott and Dudley [464]). From this approach, clinical psychologists emphasized the therapeutic benefits of working with the patients' positive experiences. Our results suggest that psychologists could work on increasing the agency of females (to be the agents of their happy moments) for improving their well-being, as agency is significantly associated with a higher level of personal satisfaction [417]. In addition, for both males and females, increasing

[6] https://www.adultdevelopmentstudy.org/.

their involvement in social contexts could benefit them in the long term [459].

We further note some of the limitations of this work. One limitation is that we are not dealing with happiness itself, but with the expression of satisfaction in natural language. We used the expression of happy moments as our work builds upon previously constructed datasets (i.e., HappyDB [398]), which provide a clean training set for training the models used in the downstream analysis. Nevertheless, the concept of happiness, in general, is difficult to define precisely, and we decided only to consider expressions of a momentary state of happiness, following other works in this area [395, 397, 398].

The RSDD [74] dataset, from which we extracted the happy moments of users in the depression and control groups, might have a demographic skew specific to Reddit towards male users.[7] We included the gender of users in the analysis, in order to avoid reaching incorrect conclusions which may have been due to gender differences, and not to depression.

Given that the majority of Reddit users come from the United States[8] and that our model for happy moments extraction was trained to detect expressions similar to HappyDB [398], which has most moments shared by people from the US, the results presented in this work may not generalize for people from other locations. To mitigate this geographical bias, novel data from different demographic groups should be collected.

Furthermore, we are relying on the RSDD data collected by [74] with users labeled by their mentions of depression diagnosis. Even if Reddit is a social media platform where people disclose their diagnosis of mental health problems and share their experiences [280], some self-mentions may not be accurate, or there may be users in the control group who suffer from depression, but chose to not talk about it online [450]. However, the significant differences we found between the two groups allow us to consider that the potential noise in the labels for the depression and control groups does not invalidate the entire dataset. The RSDD dataset still provides valuable insights into the experiences and language use of individuals suffering from depression.

Due to the lack of demographic metadata in the RSDD dataset, we used the state-of-the-art model from Tigunova et al. [403] to infer gender labels for the users. The model was trained to detect only binary gender labels, and does not consider people outside of the two traditional gender categories [450], due to data not being available for this purpose. We follow a similar approach to previous works [449–451], and consider only the male and female genders in our analyses.

[7] https://www.statista.com/statistics/1255182/distribution-of-users-on-reddit-worldwide-gender/.
[8] https://www.similarweb.com/website/reddit.com/.

## 6.8 Summary

In this chapter, we consider that the exploration of the happy moments reported by people with mentions of depression diagnoses can be used to derive insights for psychological interventions or automatic advice to initiate actions that can increase the well-being of individuals [465]. We proposed a novel approach following the Positive-Unlabeled Learning Framework for extracting happy moments from data written by individuals with mentions of depression diagnoses and control, and we evaluated it on manually annotated samples to address **RQ5**. In addition, we developed a semi-supervised approach for agency and sociality detection, two important concepts in people's expressions of happiness. We performed topic analysis and showed the differences in the topics of happy moments between users in the depression and control groups, which also vary by gender. Our results showed that males' sense of agency is not impaired because of their depression status. At the same time, females from the depression group express less agency in their expressions of happiness than control females. Furthermore, female and male users from the depression group expressed more sociality in their happy moments than control users. This research is one step more into a promising path of collaboration between computational linguistics and psychology. Analyzing the way individuals with depression express their happiness in natural language offers new pathways into the features of depression [25]. Furthermore, the results from this work could help psychologists develop new therapeutic strategies to help people suffering from depression improve their well-being.

CHAPTER 7

# Conclusions

In this final chapter, we provide the answers to the research questions introduced at the beginning of the thesis, along with our conclusions and future directions.

## 7.1 Answers to Research Questions

We further summarize the findings of the thesis by providing the answers to each of the research questions.

**RQ1 Are there significant differences in part-of-speech usage in social media texts between individuals with a self-reported depression diagnosis and a control group?**

In Section 3.1, we made a comprehensive analysis of the part-of-speech use in social media text by individuals who self-reported a depression diagnosis compared to a control group [3]. Our findings aligned with existing studies in psychology and indicated that individuals diagnosed with depression tend to use more pronouns and verbs, while using fewer nouns. We found that users with depression diagnoses frequently use first-person singular pronouns in their social media texts. This pattern reflects the self-focused attention tendency, which is consistent with psychological research. In addition, they use a smaller proportion of first-person plural pronouns, suggesting a diminished sense of belonging. In terms of verb usage, individuals with depression had a lower frequency of verbs portraying future actions, which may be linked to anhedonia—a diminished ability to experience pleasure. Interestingly, we did not find evidence of rumination—repetitive thinking about past negative events—in the social media posts of individuals with depression. These individuals used fewer past tense verbs compared to those in the control group. Another finding was a reduced use of nouns, which could indicate a diminished interest in their environment.

**RQ2 Are there significant differences in the use of offensive language between individuals with self-reported depression diagnoses and control?**

In Section 3.2, we introduced the first work that applies offensive language identification techniques to social media posts written by individuals with depression [4]. Our findings indicated that users with self-reported depression diagnoses are more likely to use offensive language in their posts compared to a control group. Among the posts from individuals with depression, those showing signs of depression contained more offensive language than those without such signs. In addition, individuals with depression used less targeted offensive language, using instead general profanity. The use of offensive language from the depression group could be an emotion regulation strategy through which depressed individuals relieve some of their distress.

**RQ3 How have recent advancements in artificial intelligence and natural language processing improved the early detection of mental disorders such as pathological gambling, self-harm, and depression, from social media texts?**

In Chapter 4, we presented the machine learning and deep learning models we employed for our participation in the early risk detection tasks organized by the eRisk Lab. We demonstrated that these models can effectively predict the risk of pathological gambling, self-harm, and depression from social media texts. Our models achieved competitive results compared to other participants in the tasks.

In our participation in the eRisk 2021 Lab (Section 4.1), we focused on the early risk detection of pathological gambling, self-harm, and estimating depression severity based on a user's social media posts [2]. The first two tasks were designed as early risk detection tasks, with the goal of making accurate predictions as soon as patterns in users' posts suggested a risk of developing these mental health disorders. For the detection of pathological gambling, due to a lack of training data, we opted for crawling and constructing a dataset from popular subreddits that discuss gambling addiction and trained a BERT classifier using these texts. Similarly, for the detection of self-harm, we gathered data from a subreddit related to self-harm behavior to augment the existing training dataset provided by the competition organizers. We then trained another BERT classifier on this dataset to predict the risk of self-harm. Our methods achieved competitive results.

In the eRisk 2022 Lab (Section 4.2), we tackled the early risk detection of pathological gambling and depression using a transformer architecture for user-level classification that is trainable end-to-end [7]. For each individual, the transformer model processed a set of texts that were encoded by a pretrained sentence encoder (RoBERTa / MiniLM) to model the interactions between posts and mitigate noise in the dataset. In addition, our model was interpretable through modern feature attribution methods and allowed

for the identification of discriminating posts in users' texts. Through our participation in this edition of the eRisk Lab, we achieved top results compared to other participants, demonstrating the effectiveness of our proposed models in early detection tasks. We achieved the best $ERDE_5$ score of 0.015, and the second-best $ERDE_{50}$ score of 0.009 for pathological gambling detection. For the early detection of depression, we obtained the second-best $ERDE_{50}$ (0.027).

In the eRisk 2023 Lab (Section 4.3), we addressed the identification of symptoms of depression in Reddit social media texts [13, 18]. The task consisted of retrieving and ranking Reddit social media sentences that convey symptoms of depression from the BDI-II questionnaire. We used synthetic data generated by GPT-3.5 and Llama-3 for each of the symptoms listed in BDI-II to augment our texts. Our aim was to enrich the data with semantic diversity and emotional and anecdotal experiences that are specific to the more intimate way of sharing experiences on Reddit. We compared the performance of two transformer-based models (MentalRoBERTa and a variant of MPNet) in embedding social media posts and the original and generated BDI-II responses. We found that using sentence embeddings from MPNet, which is designed for semantic search, outperformed the approach using embeddings from MentalRoBERTa, a model pre-trained on mental health data. Contrary to our expectations, the synthetic data did not improve the performance of the models, suggesting that data generated by GPT-3.5 and Llama-3 were too specific for retrieving symptoms of depression.

Early risk detection of mental disorders has a massive positive impact on the well-being of people, as it allows for early intervention and treatment. Through our participation in the eRisk Lab over several years, we have demonstrated that machine learning and deep learning models can effectively identify the early signs of mental disorders by analyzing social media texts.

**RQ4 Are multimodal models more effective than unimodal models in detecting depression from social media data that includes text, images, or video content?**

In Section 5.1, we introduced a flexible time-enriched multimodal transformer architecture for detecting depression from social media posts, using pretrained models for extracting image and text embeddings [13]. Our proposed models achieved state-of-the-art results on a popular multimodal Twitter dataset [163] and on *multiRedditDep* [64], which is currently the only multimodal Reddit dataset to date. We showed that a time-aware approach is suitable when the posting frequency is high, while a set-based approach is robust to dataset noise (i.e., many uninformative posts). In addition, we showed that our model is interpretable and enables the selection of the most informative posts from a user's social media timeline.

In Section 5.2, we presented the development of a multimodal temporal model

that processes multiple non-verbal cues over time to estimate depression in videos [14]. Building on psychological studies [56–58, 65], we explored different high-level modalities, including facial expression, body language, eye gaze, and voice. Our findings demonstrated improved performance in depression detection from both in-the-wild Youtube videos and recorded clinical interviews. Our model achieved state-of-the-art results on three key depression detection datasets: D-Vlog [177], DAIC-WOZ [268], and E-DAIC [270], surpassing the previous methods [173–178, 385–388]. In addition, our model is interpretable and it can provide insights into the relevance of each modality over time for individual subjects. This makes our method a promising solution for depression estimation, with potential applications in early-warning detection and suicide prevention in platforms with continuous video streams.

**RQ5 How can deep learning models identify happy moments from the social media data of individuals with self-reported depression diagnoses? Are there differences in these happy moments compared to a control group?**

In Chapter 6, we conducted a large-scale study examining happy moments expressed in social media texts of individuals who mentioned having a depression diagnosis [9, 19]. We developed a method based on the Positive-Unlabeled Framework to extract the expressions of happy moments from texts written by individuals with or without mentions of a depression diagnosis in social media. After extracting happy moments, we automatically inferred information regarding agency, sociality, topic, and gender. Our findings indicated that males' sense of agency is not impaired despite their depression status. In contrast, females in the depression group displayed less agency in their expressions of happiness compared to females in the control group. Moreover, both female and male users from the depression group exhibited a higher level of sociality in their happy moments than those in the control group. We found a common pattern among males and females in the depression group: their happy moments significantly relied on topics related to social relationships, such as friendships and romantic relationships, as well as professional development, personal memories, and self-disclosure. Analyzing the way individuals with depression express their happiness in natural language provides new insights into the features of depression. Furthermore, the results from this research could assist psychologists in developing new therapeutic strategies aimed at helping people with depression improve their well-being.

## 7.2 Ethical Considerations

Addressing ethical considerations in mental health research on social media data is crucial to ensure the privacy, confidentiality, and well-being of individuals whose data is analyzed [43, 466, 467]. In this thesis, we use previously published datasets

for the detection of mental disorders. Given their sensitive nature, all datasets were anonymized prior to use by their respective authors. All example posts included in the thesis were paraphrased to prevent the identification of individuals through online searches [466]. This approach aligns with ethical guidelines on handling sensitive data and ensures that individuals' privacy is protected throughout the research process.

In this thesis, we focused on analyzing data from individuals who have received a formal diagnosis of mental disorders. Since obtaining this information can be challenging, we primarily relied on available datasets in which users' specific mentions of mental health diagnoses from social media data were considered as a proxy for formal diagnoses. The only datasets used in this thesis that provided scores for the PHQ-8, a validated questionnaire for assessing depression, are the DAIC-WOZ dataset [268] and the E-DAIC dataset [270]. Relying on self-mentions may introduce biases. For instance, individuals who publicly disclose a formal mental health diagnosis are often more likely to have sought professional help for diagnosis and/or treatment. In addition, only certain individuals may feel comfortable sharing sensitive information about their mental health online [467]. As a result, there may be individuals within the control groups of our datasets who have lived experiences with mental disorders but do not mention this information publicly. We recognize that this specific field of study has its limitations, particularly regarding the potential inaccuracies and biases related to self-mentions of mental disorder diagnoses on social media. To overcome these challenges, future research should include clinical assessments or standardized interviews to improve the validity of the findings.

We acknowledge that the methods proposed in this thesis have been trained on data with demographic bias [468] from Reddit and Twitter, two social media platforms with a demographic skew towards young males from the United States; thus, our methods may not succeed in identifying depression from other demographic groups. Cultural factors influence how distress is perceived, expressed, interpreted, and addressed, highlighting the need for a nuanced approach in the application of algorithmic abstractions [469]. In Western societies, individuals often share their emotions openly, whereas in Asian cultures, emotional struggles are frequently internalized or expressed indirectly due to collectivist values [470, 471]. For instance, self-criticism, typically associated with depression, can be seen as healthy functioning in East Asian contexts [472]. Cultural interpretations of mental health symptoms can lead some individuals to report somatic complaints instead of psychological distress [469]. An example is the term "nervios," commonly found in Latin American communities, which manifests with psychological and somatic symptoms and has a high comorbidity with anxiety and mood disorders [473]. The DSM-V [30] includes cultural concepts of distress to aid clinicians in recognizing diverse expressions of mental health issues. In addition, previous studies have shown that English-speaking individuals from Western countries often use more clinical language compared to users from countries like India or Malaysia [474]. Metaphors are frequently used to convey experiences of mental illness, with depression often

depicted using ideas like weight or darkness [475]. Understanding cultural and linguistic variations is crucial when developing automated methods for predicting mental disorders, as they help explain why many predictive models struggle to generalize effectively on data from different demographics [23]. Furthermore, each individual's experience with depression is unique, and it is important to consider their distinct experiences and symptomatology. Algorithmic representations and abstractions play a crucial role in the understanding of mental illness and well-being by providing a framework for generalization [476]. While these simplifications can help identify trends and better understand complex individual experiences, they also risk oversimplifying those experiences. It's important to recognize that generalizing can sometimes lead to misunderstandings regarding the unique nuances of mental health experiences and symptoms. Each person's experience with mental health disorders is unique, and acknowledging this is essential for a deeper understanding of mental health.

The methods proposed in Chapters 4 and 5 have shown promising results in detecting mental disorders through textual, visual, and non-verbal patterns. However, it is crucial to emphasize that these methods should not be used for clinical diagnosis, which must be conducted by a qualified mental health professional. While our approach can be beneficial for initial depression screening, it should only be implemented in a human-in-the-loop scenario. We prioritized the interpretability of our methods and have developed models that allow for an understanding of the predictions made. This means that experts reviewing our model's predictions can gain insights into the patterns that influenced specific decisions. By doing so, they can better understand the factors driving each prediction. Incorrect predictions can have harmful effects on individuals' lives. For example, if a system falsely predicts that someone shows signs of mental disorders, it can adversely impact their well-being due to the stigma associated with such labels. This may lead individuals to believe there is something wrong with them, ultimately lowering their self-esteem [467]. A false negative prediction occurs when the system fails to identify significant signs of distress, preventing the individual from receiving the necessary treatment or interventions. False negative predictions are particularly critical in cases of suicidal ideation, where a person's life may be at risk.

Inferring the gender of online users poses significant ethical and methodological challenges. Gender inference from text relies on assumptions that linguistic patterns directly correlate with gender and that individuals consistently express their gender through language, which may lead to misclassification and overlook the diverse nature of gender identity. Moreover, individuals do not always conform to traditional gender categories. Despite these limitations, in Chapter 6, we adopted the prevailing methodology from psychology, which commonly uses binary gender categories [425, 477, 478], to align with existing literature and facilitate comparative analyses.

## 7.3 Conclusions and Future Directions

In conclusion, this thesis explored computational methods for detecting depression, self-harm, and pathological gambling using social media data. We first highlighted the differences in part-of-speech and offensive language use in social media texts between individuals with self-reported depression diagnoses and control groups. In addition, our participation in the eRisk Lab demonstrated how effective computational approaches can be for the early detection of risks related to depression, self-harm, and pathological gambling. We have shifted from relying solely on textual data to utilizing multimodal data and proposed two architectures to process multimodal information. The first architecture processed multimodal posts that include text and images from social media, and achieved state-of-the-art results in detecting depression. Furthermore, we developed a multimodal temporal model that analyzes various non-verbal cues over time to assess depression in videos. This thesis also included a comprehensive examination of the happy moments shared by individuals with self-mentions of a depression diagnosis. We hope that our findings can aid psychologists in formulating new therapeutic strategies to help individuals with depression enhance their well-being.

In this thesis, we focused solely on using LLMs to generate synthetic data to identify symptoms of depression. However, we anticipate that LLMs will play a significant role in future research, especially since these models can offer more comprehensive explanations for their decisions [15, 479]. In addition, we see future directions involving the evaluation of LLMs in multilingual settings and the development of computational approaches specifically designed for detecting mental disorders from multilingual data [23].

It is important to address the limitations surrounding the potential inaccuracies and biases associated with self-mentions of mental disorder diagnoses on social media. To overcome these challenges, future research should incorporate clinical assessments or standardized interviews for mental disorders assessment to enhance the validity of the findings. Furthermore, avoiding reliance on binary labels when representing diagnoses can help capture the nuances of the spectrum of mental illness severity [476]. In future work, we aim to consider the unique aspects of individual mental health experiences and symptoms, as each person's experience with mental health disorders is distinct.

# Bibliography


[1] **A.-M. Bucur** and L. P. Dinu, *Detecting Early Onset of Depression From Social Media Text Using Learned Confidence Scores*, Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it) (2020) (cit. on pp. xv, 42, 58).

[2] **A.-M. Bucur**, A. Cosma and L. P. Dinu, *Early Risk Detection of Pathological Gambling, Self-Harm and Depression Using BERT*, Working Notes of CLEF 2021, Early Risk Prediction on the Internet Lab (eRisk), CEUR 2936 (2021) pp. 938–949 (cit. on pp. xv, 4, 5, 7, 14, 16, 51, 52, 57, 59, 62, 71, 73, 101, 125).

[3] **A.-M. Bucur**, I. R. Podina and L. P. Dinu, *A Psychologically Informed Part-of-Speech Analysis of Depression in Social Media*, Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP) (2021) pp. 199–207 (cit. on pp. xv, 4, 6, 9, 14, 37, 39, 49, 124).

[4] **A.-M. Bucur**, M. Zampieri and L. P. Dinu, *An Exploratory Analysis of the Relation Between Offensive Language and Mental Health*, Findings of the Association for Computational Linguistics: ACL-IJCNLP (2021) pp. 3600–3606 (cit. on pp. xv, 4, 6, 14, 37, 47, 50, 125).

[5] **A.-M. Bucur**, A. Cosma and L. P. Dinu, *Sequence-to-Sequence Lexical Normalization With Multilingual Transformers*, Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT), EMNLP (2021) pp. 473–482 (cit. on pp. xv, 38).

[6] **A.-M. Bucur**, H. Jang and F. F. Liza, *Capturing Changes in Mood Over Time in Longitudinal Data Using Ensemble Methodologies*, Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology (CLPsych), NAACL (2022) pp. 205–212 (cit. on pp. xv, 13).

[7] **A.-M. Bucur**, A. Cosma, L. P. Dinu and P. Rosso, *An End-to-End Set Transformer for User-Level Classification of Depression and Gambling Disorder*, Working Notes of CLEF 2022, Early Risk

Prediction on the Internet Lab (eRisk), CEUR 3180 (2022) pp. 851–863 (cit. on pp. xv, 4, 7, 14, 51, 57, 72, 125).

[8] **A.-M. Bucur**, A. Cosma and I.-B. Iordache, *BLUE at Memotion 2.0 2022: You Have My Image, My Text and My Transformer*, Proceedings of the First Workshop on Multimodal Fact-Checking and Hate Speech Detection (DE-FACTIFY), AAAI (2022) (cit. on pp. xv, 78).

[9] **A.-M. Bucur**, A. Cosma and L. P. Dinu, *Life Is Not Always Depressing: Exploring the Happy Moments of People Diagnosed With Depression*, Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC) (2022) pp. 4186–4192 (cit. on pp. xvi, 6, 8, 98, 102, 127).

[10] **A.-M. Bucur**, *Utilizing ChatGPT Generated Data to Retrieve Depression Symptoms From Social Media*, Working Notes of CLEF 2023, Early Risk Prediction on the Internet Lab (eRisk), CEUR 3497 (2023) pp. 662–671 (cit. on pp. xvi, 4, 7, 15, 51, 65, 72).

[11] I. R. Podina et al., *Mental Health at Different Stages of Cancer Survival: A Natural Language Processing Study of Reddit Posts*, Frontiers in Psychology 14 (2023) pp. 1150227 (cit. on pp. xvi, 14).

[12] I. R. Podina, **A.-M. Bucur**, L. Fodor and R. Boian, *Screening for Common Mental Health Disorders: A Psychometric Evaluation of a Chatbot System*, Behaviour & Information Technology (2023) (cit. on pp. xvi, 1).

[13] **A.-M. Bucur**, A. Cosma, P. Rosso and L. P. Dinu, *It's Just a Matter of Time: Detecting Depression With Time-Enriched Multimodal Transformers*, Advances in Information Retrieval. ECIR 2023. Lecture Notes in Computer Science 13980 (2023) pp. 200–215 (cit. on pp. xvi, 5, 7, 17, 73, 75, 85, 97, 101, 126).

[14] D. Gimeno-Gómez, **A.-M. Bucur**, A. Cosma, C.-D. Martínez-Hinarejos and P. Rosso, *Reading Between the Frames: Multi-Modal Depression Detection in Videos From Non-Verbal Cues*, Advances in Information Retrieval. ECIR 2024. Lecture Notes in Computer Science 14608 (2024) pp. 191–209 (cit. on pp. xvi, 5, 7, 17, 73, 97, 127).

[15] V. Varadarajan et al., *Archetypes and Entropy: Theory-Driven Extraction of Evidence for Suicide Risk*, Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych), EACL (2024) pp. 278–291 (cit. on pp. xvi, 12, 130).

[16] M. N. Raihan et al., *MentalHelp: A Multi-Task Dataset for Mental Health in Social Media*, Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING) (2024) pp. 11196–11203 (cit. on pp. xvi, 15, 24).

[17] R. Boian et al., *A Conversational Agent Framework for Mental Health Screening: Design, Implementation, and Usability*, Behaviour & Information Technology (2024) (cit. on pp. xvi, 1).

[18] **A.-M. Bucur**, *Leveraging LLM-Generated Data for Detecting Depression Symptoms on Social Media*, Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2024. Lecture Notes in Computer Science 14958 (2024) pp. 193–204 (cit. on pp. xvii, 4, 7, 51, 65, 72, 126).

[19] **A.-M. Bucur**, B. Chulvi, A. Cosma and P. Rosso, *The Expression of Happiness in Social Media of Individuals Reporting Depression*, IEEE Transactions on Affective Computing 16 (2025) pp. 360–375 (cit. on pp. xvii, 6, 8, 98, 101, 127).

[20] **A.-M. Bucur** et al., *On the State of NLP Approaches to Modeling Depression in Social Media: A Post-Covid-19 Outlook*, IEEE Journal of Biomedical and Health Informatics (2025) (cit. on pp. xvii, 25).

[21] **A.-M. Bucur** et al., *Datasets for Depression Modeling in Social Media: An Overview*, Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych), NAACL (2025) (cit. on pp. xvii, 18).

[22] A. Sandu, T. Mihailescu, A. S. Uban and **A.-M. Bucur**, *Capturing the Dynamics of Mental Well-Being: Adaptive and Maladaptive States in Social Media*, Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych), NAACL (2025) (cit. on pp. xvii, 12).

[23] **A.-M. Bucur**, M. Zampieri, T. Ranasinghe and F. Crestani, *A Survey on Multilingual Mental Disorders Detection From Social Media Data*, Pre-Print (2025) (cit. on pp. xvii, 129, 130).

[24] N. Raihan, **A.-M. Bucur** and M. Zampieri, *Large Language Models for Mental Health: A Multilingual Evaluation*, Pre-Print (2025) (cit. on pp. xvii, 15).

[25] **A.-M. Bucur** et al., *Vanishing Boundaries: A Unifying Account of Multidimensional Emotion Dynamics and Alterations in Depression*, Pre-Print (2025) (cit. on pp. xvii, 123).

[26] J. J. McGrath et al., *Age of Onset and Cumulative Risk of Mental Disorders: A Cross-National Analysis of Population Surveys From 29 Countries*, The Lancet Psychiatry 10 (2023) pp. 668–681 (cit. on p. 1).

[27] A. A. Ahad, M. Sanchez-Gonzalez and P. Junquera, *Understanding and Addressing Mental Health Stigma Across Cultures for Improving Psychiatric Care: A Narrative Review*, Cureus 15 (2023) (cit. on p. 1).

[28] R. R. Bond et al., *Digital Transformation of Mental Health Services*, Npj Mental Health Research 2 (2023) pp. 13 (cit. on p. 1).

[29] A. T. Beck et al., *Cognitive therapy of depression*, (1979) (cit. on p. 1).

[30] A. P. Association et al., *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*, 5th ed., Washington, DC: American Psychiatric Pub, 2013 (cit. on pp. 1, 2, 10, 12, 40, 52, 85, 128).

[31] M. Ratcliffe, *Varieties of temporal experience in depression*, Journal of Medicine and Philosophy 37 (2012) pp. 114–138 (cit. on p. 2).

[32] W. Styron, *Darkness visible: A memoir of madness*, (1990) (cit. on p. 2).

[33] A. Cipriano, S. Cella and P. Cotrufo, *Nonsuicidal Self-Injury: A Systematic Review*, Frontiers in Psychology 8 (2017) pp. 1946 (cit. on p. 2).

[34] D. Gillies et al., *Prevalence and Characteristics of Self-Harm in Adolescents: Meta-Analyses of Community-Based Studies 1990–2015*, Journal of the American Academy of Child & Adolescent Psychiatry 57 (2018) pp. 733–741 (cit. on p. 2).

[35] S. Gainsbury and R. Wood, *Internet Gambling Policy in Critical Comparative Perspective: The Effectiveness of Existing Regulatory Frameworks*, International Gambling Studies 11 (2011) pp. 309–323 (cit. on pp. 2, 52).

[36] S. M. Gainsbury, *Online Gambling Addiction: The Relationship Between Internet Gambling and Disordered Gambling*, Current Addiction Reports 2 (2015) pp. 185–193 (cit. on pp. 2, 52).

[37] S. Rude, E.-M. Gortner and J. Pennebaker, *Language Use of Depressed and Depression-Vulnerable College Students*, Cognition & Emotion 18 (2004) pp. 1121–1133 (cit. on pp. 2, 3, 9, 49).

[38] S. Fekete, *The Internet - A New Source of Data on Suicide, Depression and Anxiety: A Preliminary Study*, Archives of Suicide Research 6 (2002) pp. 351–361 (cit. on pp. 2, 3, 9).

[39] E. C. Stade, L. Ungar, J. C. Eichstaedt, G. Sherman and A. M. Ruscio, *Depression and Anxiety Have Distinct and Overlapping Language Patterns: Results From a Clinical Interview.*, Journal of Psychopathology and Clinical Science (2023) (cit. on pp. 2, 9).

[40] M. De Choudhury, M. Gamon, S. Counts and E. Horvitz, *Predicting Depression via Social Media*, Proceedings of the International AAAI Conference on Web and Social Media (2013) (cit. on pp. 2, 9, 17, 79).

[41] J. C. Eichstaedt et al.,
*Facebook Language Predicts Depression in Medical Records*, Proceedings of the National Academy of Sciences 115 (2018) pp. 11203–11208 (cit. on pp. 2, 14, 21).

[42] G. Coppersmith, K. Ngo, R. Leary and A. Wood,
*Exploratory Analysis of Social Media Prior to a Suicide Attempt*,
Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology (2016) pp. 106–117 (cit. on p. 2).

[43] S. Chancellor and M. De Choudhury, *Methods in Predictive Techniques for Mental Health Status on Social Media: A Critical Review*,
NPJ digital medicine 3 (2020) pp. 1–11 (cit. on pp. 3, 127).

[44] E. A. Ríssola, D. E. Losada and F. Crestani, *A Survey of Computational Methods for Online Mental State Assessment on Social Media*,
ACM Transactions on Computing for Healthcare 2 (2021) pp. 1–31 (cit. on p. 3).

[45] S. Amir, M. Dredze and J. W. Ayers,
*Mental Health Surveillance Over Social Media With Digital Cohorts*,
Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology (2019) pp. 114–120 (cit. on pp. 3, 19).

[46] R. Skaik and D. Inkpen, *Using Twitter Social Media for Depression Detection in the Canadian Population*,
Proceedings of the 2020 3rd Artificial Intelligence and Cloud Computing Conference (2020) pp. 109–114 (cit. on pp. 3, 16).

[47] D. Smirnova et al., *Language Patterns Discriminate Mild Depression From Normal Sadness and Euthymic State*,
Frontiers in Psychiatry 9 (2018) pp. 105 (cit. on pp. 3, 9, 40).

[48] M. L. Birnbaum et al., *Identifying Signals Associated With Psychiatric Illness Utilizing Language and Images Posted to Facebook*,
npj Schizophrenia 6 (2020) pp. 1–10 (cit. on pp. 4, 14, 45, 85).

[49] H. Alhuzali, T. Zhang and S. Ananiadou, *Predicting Sign of Depression via Using Frozen Pre-Trained Models and Random Forest Classifier*,
Working Notes of CLEF 2021, Early Risk Prediction on the Internet Lab (eRisk), CEUR 2936 (2021) pp. 888–896 (cit. on pp. 5, 16, 73, 74).

[50] S.-H. Wu and Z.-J. Qiu, *A RoBERTa-based Model on Measuring the Severity of the Signs of Depression*,
Working Notes of CLEF 2021, Early Risk Prediction on the Internet Lab (eRisk), CEUR 2936 (2021) pp. 1071–1080 (cit. on pp. 5, 14, 16, 73, 74).

[51] P. Mann, A. Paes and E. H. Matsushima,
*See and Read: Detecting Depression Symptoms in Higher Education Students Using Multimodal Social Media Data*,
Proceedings of the International AAAI Conference on Web and social media 14 (2020) pp. 440–451 (cit. on pp. 5, 17, 73, 74).

[52] R. Safa, P. Bayat and L. Moghtader, *Automatic Detection of Depression Symptoms in Twitter Using Multimodal Analysis*,
The Journal of Supercomputing 78 (2022) (cit. on pp. 5, 17, 73, 74).

[53] Z. Xu, V. Pérez-Rosas and R. Mihalcea, *Inferring Social Media Users' Mental Health Status From Multimodal Information*,
Proceedings of the Twelfth Language Resources and Evaluation Conference (2020) pp. 6292–6299 (cit. on pp. 5, 17, 73, 74).

[54] A. G. Reece and C. M. Danforth,
*Instagram Photos Reveal Predictive Markers of Depression*,
EPJ Data Science 6 (2017) pp. 15 (cit. on pp. 5, 10, 21, 85).

[55] S. C. Guntuku, D. Preotiuc-Pietro, J. C. Eichstaedt and L. H. Ungar, *What Twitter Profile and Posted Images Reveal About Depression and Anxiety*,
Proceedings of the International AAAI Conference on Web and Social Media 13 (2019) pp. 236–246 (cit. on pp. 5, 10, 22).

[56] B. Renneberg, K. Heyn, R. Gebhard and S. Bachmann, *Facial Expression of Emotions in Borderline Personality Disorder and Depression*, Journal of Behavior Therapy and Experimental Psychiatry 36 (2005) pp. 183–196 (cit. on pp. 5, 10, 85, 88, 97, 127).

[57] M. Yamamoto et al.,
*Using Speech Recognition Technology to Investigate the Association Between Timing-Related Speech Features and Depression Severity*,
PloS One 15 (2020) e0238726 (cit. on pp. 5, 10, 85, 89, 97, 127).

[58] J. T. Fiquer, P. S. Boggio and C. Gorenstein, *Talking Bodies: Nonverbal Behavior in the Assessment of Depression Severity*,
Journal of Affective Disorders 150 (2013) pp. 1114–1119
(cit. on pp. 5, 10, 85, 89, 97, 127).

[59] M. E. Beutel, H. Glaesmer, J. Wiltink, H. Marian and E. Brähler,
*Life Satisfaction, Anxiety, Depression and Resilience Across the Life Span of Men*, The Aging Male 13 (2010) pp. 32–39 (cit. on pp. 5, 100).

[60] E. H. Seo et al.,
*Life Satisfaction and Happiness Associated With Depressive Symptoms Among University Students: A Cross-Sectional Study in Korea*,
Annals of General Psychiatry 17 (2018) pp. 1–9 (cit. on pp. 5, 100).

[61] A. Bergsma, M. T. Have, R. Veenhoven and R. D. Graaf, *Most People With Mental Disorders Are Happy: A 3-Year Follow-Up in the Dutch General Population*, Journal of Positive Psychology 6 (2011) pp. 253–259 (cit. on pp. 5, 100, 101).

[62] C. E. Osgood and E. G. Walker, *Motivation and Language Behavior: A Content Analysis of Suicide Notes.*, The Journal of Abnormal and Social Psychology 59 (1959) pp. 58 (cit. on p. 9).

[63] R. A. Sansone and L. A. Sansone, *Rumination: Relationships With Physical Health*, Innovations in Clinical Neuroscience 9 (2012) pp. 29 (cit. on pp. 9, 40).

[64] A.-S. Uban, B. Chulvi and P. Rosso, *Explainability of Depression Detection on Social Media: From Deep Learning Models to Psychological Interpretations and Multimodality*, Early Detection of Mental Health Disorders by Social Media Monitoring (2022) pp. 289–320 (cit. on pp. 10, 23, 35, 74, 77–82, 97, 126).

[65] D. Bennabi, P. Vandel, C. Papaxanthis, T. Pozzo and E. Haffen, *Psychomotor Retardation in Depression: A Systematic Review of Diagnostic, Pathophysiologic, and Therapeutic Implications*, BioMed Research International 2013 (2013) (cit. on pp. 10, 85, 88, 97, 127).

[66] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. Mavadati and D. P. Rosenwald, *Social Risk and Depression: Evidence From Manual and Automatic Facial Expression Analysis*, 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) (2013) pp. 1–8 (cit. on pp. 10, 88).

[67] Y. Zhang et al., *Associations Between Depression Symptom Severity and Daily-Life Gait Characteristics Derived From Long-Term Acceleration Signals in Real-World Settings: Retrospective Analysis*, JMIR mHealth and uHealth 10 (2022) e40667 (cit. on pp. 10, 89).

[68] J. Rottenberg and C. Vaughan, *Emotion Expression in Depression: Emerging Evidence for Emotion Context-Insensitivity*, Emotion Regulation: Conceptual and Clinical Issues (2008) pp. 125–139 (cit. on pp. 10, 85, 89).

[69] A. Pavlidou et al., *Hand Gesture Performance Is Impaired in Major Depressive Disorder: A Matter of Working Memory Performance?*, Journal of Affective Disorders 292 (2021) pp. 81–88 (cit. on pp. 10, 89).

[70] J. Wang et al., *Acoustic Differences Between Healthy and Depressed People: A Cross-Situation Study*, BMC Psychiatry 19 (2019) pp. 1–12 (cit. on pp. 10, 89).

[71] J. Mackintosh, R. Kumar and T. Kitamura, *Blink Rate in Psychiatric Illness*, The British Journal of Psychiatry 143 (1983) pp. 55–57 (cit. on pp. 10, 85, 89).

[72] G. M. Lucas, J. Gratch, S. Scherer, J. Boberg and G. Stratou, *Towards an Affective Interface for Assessment of Psychological Distress*, 2015 International Conference on Affective Computing and Intelligent Interaction (ACII) (2015) pp. 539–545 (cit. on pp. 10, 85, 89).

[73] B. P. Taylor et al., *Psychomotor Slowing as a Predictor of Fluoxetine Nonresponse in Depressed Outpatients*, American Journal of Psychiatry 163 (2006) pp. 73–78 (cit. on pp. 10, 85).

[74] A. Yates, A. Cohan and N. Goharian, *Depression and Self-Harm Risk Assessment in Online Forums*, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (2017) pp. 2968–2978 (cit. on pp. 10, 11, 14, 16, 18, 21, 34, 46, 53, 74, 86, 101, 102, 104, 105, 122).

[75] A. Benton, M. Mitchell and D. Hovy, *Multitask Learning for Mental Health Conditions With Limited Social Media Data*, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers (2017) pp. 152–162 (cit. on pp. 11, 21).

[76] G. Shen et al., *Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution*, Proceedings of the 26th International Joint Conference on Artificial Intelligence (2017) pp. 3838–3844 (cit. on pp. 11, 14, 17, 18, 21, 32, 35, 74, 77).

[77] D. E. Losada, F. Crestani and J. Parapar, *CLEF 2017 eRisk Overview: Early Risk Prediction on the Internet: Experimental Foundations.*, Working Notes of CLEF 2017, Early Risk Prediction on the Internet Lab (eRisk), CEUR 1866 (2017) (cit. on pp. 11, 17, 21).

[78] R. A. Calvo, D. N. Milne, M. S. Hussain and H. Christensen, *Natural Language Processing in Mental Health Applications Using Non-Clinical Texts*, Natural Language Engineering 23 (2017) pp. 649–685 (cit. on p. 11).

[79] D. Lee, S. Park, J. Kang, D. Choi and J. Han, *Cross-Lingual Suicidal-Oriented Word Embedding Toward Suicide Prevention*, Findings of the Association for Computational Linguistics: EMNLP 2020 (2020) pp. 2208–2217 (cit. on p. 11).

[80] A. Haque, V. Reddi and T. Giallanza, *Deep Learning for Suicide and Depression Identification With Unsupervised Label Correction*, Artificial Neural Networks and Machine Learning – ICANN 2021 (2021) pp. 436–447, ed. by I. Farkaš, P. Masulli, S. Otte and S. Wermter (cit. on pp. 11, 14, 22).

[81] I. Pirina and Ç. Çöltekin, *Identifying Depression on Reddit: The Effect of Training Data*, Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop and Shared Task (2018) pp. 9–12 (cit. on pp. 11, 18, 21).

[82] D. Owen, J. C. Collados and L. Espinosa-Anke, *Towards Preemptive Detection of Depression and Anxiety in Twitter*, Proceedings of the Fifth Social Media Mining for Health Applications Workshop and Shared Task (2020) pp. 82–89 (cit. on pp. 11, 22).

[83] J. Wolohan, M. Hiraga, A. Mukherjee, Z. A. Sayyed and M. Millard, *Detecting Linguistic Traces of Depression in Topic-Restricted Text: Attending to Self-Stigmatized Depression With NLP*, Proceedings of the First International Workshop on Language Cognition and Computational Models (2018) pp. 11–21 (cit. on pp. 11, 21, 27, 86).

[84] A. H. Yazdavar et al., *Multimodal Mental Health Analysis in Social Media*, PLoS One 15 (2020) (cit. on pp. 11, 14, 22).

[85] D. Zarate, M. Ball, M. Prokofieva, V. Kostakos and V. Stavropoulos, *Identifying Self-Disclosed Anxiety on Twitter: A Natural Language Processing Approach*, Psychiatry Research 330 (2023) pp. 115579 (cit. on p. 11).

[86] I. Sekulić, M. Gjurković and J. Šnajder, *Not Just Depressed: Bipolar Disorder Prediction on Reddit*, Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (2018) pp. 72–78 (cit. on p. 11).

[87] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead and M. Mitchell, *CLPsych 2015 Shared Task: Depression and PTSD on Twitter*, Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (2015) pp. 31–39 (cit. on pp. 11, 13, 18).

[88] J. Parapar, P. Martin-Rodilla, D. E. Losada and F. Crestani, *Overview of eRisk 2021: Early Risk Prediction on the Internet.*, Working Notes of CLEF 2021, Early Risk Prediction on the Internet Lab (eRisk), CEUR 2936 (2021) pp. 864–887 (cit. on pp. 12, 17, 22, 34).

[89] J. Parapar, P. Martín-Rodilla, D. E. Losada and F. Crestani, *Overview of eRisk 2023: Early Risk Prediction on the Internet*, Experimental IR Meets Multilinguality, Multimodality, and Interaction. 14th International Conference of the CLEF Association (2023) (cit. on pp. 12, 13, 34, 65).

[90] U. Naseem, A. G. Dunn, J. Kim and M. Khushi, *Early Identification of Depression Severity Levels on Reddit Using Ordinal Classification*, Proceedings of the ACM Web Conference, WWW '22 (2022) pp. 2563–2572 (cit. on pp. 12, 22).

[91] M. Kabir et al., *DEPTWEET: A Typology for Social Media Texts to Detect Depression Severities*, Computers in Human Behavior 139 (2023) pp. 107503 (cit. on pp. 12, 23).

[92] K. Sampath and T. Durairaj, *Data Set Creation and Empirical Analysis for Detecting Signs of Depression From Social Media Postings*, Computational Intelligence in Data Science (2022) pp. 136–151 (cit. on pp. 12, 17, 23).

[93] A. Zirikly, P. Resnik, O. Uzuner and K. Hollingshead, *CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts*, Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology (2019) pp. 24–33 (cit. on pp. 12, 13).

[94] M. Gaur et al., *Knowledge-Aware Assessment of Severity of Suicide Risk for Early Intervention*, Proceedings of the World Wide Web Conference (2019) pp. 514–525 (cit. on p. 12).

[95] K. Posner et al., *Columbia-Suicide Severity Rating Scale (C-Ssrs)*, New York, NY: Columbia University Medical Center 10 (2008) pp. 2008 (cit. on p. 12).

[96] X. Xu et al., *Leveraging Large Language Models for Mental Health Prediction via Online Text Data*, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (2023) (cit. on pp. 12, 15).

[97] K. Yang et al., *Towards Interpretable Mental Health Analysis With Large Language Models*, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (2023) pp. 6056–6077 (cit. on pp. 12, 15).

[98] W. Qin et al., *Read, Diagnose and Chat: Towards Explainable and Interactive LLMs-Augmented Depression Detection in Social Media*, arXiv preprint arXiv:2305.05138 (2023) (cit. on pp. 12, 15).

[99] M. M. Amin, E. Cambria and B. W. Schuller, *Will Affective Computing Emerge From Foundation Models and General AI? A First Evaluation on ChatGPT*, IEEE Intelligent Systems 38 (2023) pp. 2 (cit. on pp. 12, 15).

[100] A. Pérez, M. Fernández-Pichel, J. Parapar and D. E. Losada, *DepreSym: A Depression Symptom Annotated Corpus and the Role of LLMs as Assessors of Psychological Markers*, arXiv preprint arXiv:2308.10758 (2023) (cit. on pp. 12, 15, 23).

[101] K. Kroenke, R. L. Spitzer and J. B. Williams, *The PHQ-9: Validity of a Brief Depression Severity Measure*, Journal of General Internal Medicine 16 (2001) pp. 606–613 (cit. on p. 12).

[102] A. T. Beck, R. A. Steer and G. Brown, *Beck Depression Inventory–Ii*, Psychological Assessment (1996) (cit. on pp. 12, 34, 65).

[103] J. Chim et al., *Overview of the Clpsych 2024 Shared Task: Leveraging Large Language Models to Identify Evidence of Suicidality Risk in Online Posts*, Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych) (2024) pp. 177–190 (cit. on p. 12).

[104] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith and M. Kumar, *Discovering Shifts to Suicidal Ideation From Mental Health Content in Social Media*, Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (2016) pp. 2098–2110 (cit. on pp. 13, 39, 42).

[105] A. Tsakalidis et al., *Overview of the CLPsych 2022 Shared Task: Capturing Moments of Change in Longitudinal User Posts*, Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology (2022) pp. 184–198 (cit. on p. 13).

[106] A. Tsakalidis et al., *Identifying Moments of Change From Longitudinal User Text*, Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2022) pp. 4647–4660 (cit. on pp. 13, 77).

[107] C. Cohrdes et al., *Indications of Depressive Symptoms During the COVID-19 Pandemic in Germany: Comparison of National Survey and Twitter Data*, JMIR Mental Health 8 (2021) e27140 (cit. on pp. 13, 26, 28, 29).

[108] D. Losada and F. Crestani, *A Test Collection for Research on Depression and Language Use*, Experimental IR Meets Multilinguality, Multimodality, and Interaction, 7th International Conference of the CLEF Association, CLEF (2016) pp. 28–39 (cit. on pp. 13, 53, 62).

[109] V. Lynn et al., *CLPsych 2018 Shared Task: Predicting Current and Future Psychological Health From Childhood Essays*, Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic (2018) pp. 37–46 (cit. on p. 13).

[110] S. Kayalvizhi, T. Durairaj, B. R. Chakravarthi et al., *Findings of the Shared Task on Detecting Signs of Depression From Social Media*, Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion (2022) pp. 331–338 (cit. on p. 13).

[111] S. Kayalvizhi et al., *Overview of the Shared Task on Detecting Signs of Depression From Social Media Text*, Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion (2023) pp. 25–30 (cit. on p. 13).

[112] A. M. Mármol-Romero et al., *Overview of MentalriskES at IberLEF 2023: Early Detection of Mental Disorders Risk in Spanish*, Procesamiento del Lenguaje Natural 71 (2023) pp. 329–350 (cit. on p. 13).

[113] A. Cohan et al., *SMHD: A Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions*, Proceedings of the 27th International Conference on Computational Linguistics (2018) pp. 1485–1497 (cit. on pp. 14, 18, 21).

[114] Z. Jamil, D. Inkpen, P. Buddhitha and K. White, *Monitoring Tweets for Depression to Detect at-Risk Users*, Proceedings of the Workshop on Computational Linguistics and Clinical Psychology (CLPsych), NAACL (2017) pp. 32–40 (cit. on pp. 14, 21).

[115] H. S. Alsagri and M. Ykhlef, *Machine Learning-Based Approach for Depression Detection in Twitter Using Content and Activity Features*, IEICE Transactions on Information and Systems E103.D (2020) pp. 1825–1832 (cit. on p. 14).

[116] Z. Peng, Q. Hu and J. Dang, *Multi-Kernel SVM Based Depression Recognition Using Social Media Data*, International Journal of Machine Learning and Cybernetics 10 (2019) pp. 43–57 (cit. on p. 14).

[117] G. Li, B. Li, L. Huang, S. Hou et al., *Automatic Construction of a Depression-Domain Lexicon Based on Microblogs: Text Mining Study*, JMIR Medical Informatics 8 (2020) e17650 (cit. on p. 14).

[118] N. Vedula and S. Parthasarathy, *Emotional and Linguistic Cues of Depression From Social Media*, Proceedings of the 2017 International Conference on Digital Health (2017) pp. 127–136 (cit. on pp. 14, 21).

[119] F. Ramiandrisoa, J. Mothe, F. Benamara and V. Moriceau, *IRIT at E-Risk 2018*, Working Notes of CLEF 2018, Early Risk Prediction on the Internet Lab (eRisk), CEUR 2125 (2018) (cit. on p. 14).

[120] A.-S. Uban and P. Rosso, *Deep Learning Architectures and Strategies for Early Detection of Self-Harm and Depression Level Prediction*, Working Notes of CLEF 2020, Early Risk Prediction on the Internet Lab (eRisk), CEUR 2696 (2020) pp. 1–12 (cit. on pp. 14, 53, 101).

[121] J. W. Pennebaker, M. E. Francis and R. J. Booth, *Linguistic Inquiry and Word Count: LIWC 2001*, Mahway: Lawrence Erlbaum Associates 71 (2001) pp. 2001 (cit. on pp. 14, 58).

[122] K. Loveys, J. Torrez, A. Fine, G. Moriarty and G. Coppersmith, *Cross-Cultural Differences in Language Markers of Depression Online*, Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic (2018) pp. 78–87 (cit. on pp. 14, 21).

[123] T. Nalabandian and M. Ireland, *Depressed Individuals Use Negative Self-Focused Language When Recalling Recent Interactions With Close Romantic Partners but Not Family or Friends*, Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology (2019) pp. 62–73 (cit. on p. 14).

[124] P. Resnik, A. Garron and R. Resnik, *Using Topic Modeling to Improve Prediction of Neuroticism and Depression in College Students*, Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (2013) pp. 1348–1353 (cit. on p. 14).

[125] S. W. Stirman and J. W. Pennebaker, *Word Use in the Poetry of Suicidal and Nonsuicidal Poets*, Psychosomatic Medicine 63 (2001) pp. 517–522 (cit. on p. 14).

[126] I. Goodfellow, *Deep Learning*, vol. 196, MIT press, 2016 (cit. on p. 14).

[127] Z. Hamad, R. Imran, M. S. Jameel and X. Guandong, *DepressionNet: A Novel Summarization Boosted Deep Framework for Depression Detection on Social Media*, Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (2021) pp. 133–142 (cit. on p. 14).

[128] C. Y. Chiu, H. Y. Lane, J. L. Koh and A. L. Chen, *Multimodal Depression Detection on Instagram Considering Time Interval of Posts*, Journal of Intelligent Information Systems 56 (2021) pp. 25–47 (cit. on pp. 14, 17, 22, 74).

[129] Y. Suhara, Y. Xu and A. Pentland, *Deepmood: Forecasting Depressed Mood Based on Self-Reported Histories via Recurrent Neural Networks*, Proceedings of ACM Web Conference (2017) pp. 715–724 (cit. on p. 14).

[130] A. H. Orabi, P. Buddhitha, M. H. Orabi and D. Inkpen, *Deep Learning for Depression Detection of Twitter Users*, Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic (2018) pp. 88–97 (cit. on p. 14).

[131] F. Sadeque, D. Xu and S. Bethard,
*Measuring the Latency of Depression Detection in Social Media*,
Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (2018) pp. 495–503 (cit. on pp. 14, 53, 62).

[132] I. A. N. Arachchige, V. H. Jayasuriya and R. Weerasinghe, *A Dataset for Research on Modelling Depression Severity in Online Forum Data*,
Proceedings of the Student Research Workshop Associated with RANLP 2021 (2021) pp. 144–153 (cit. on pp. 14, 22).

[133] A. Wołk, K. Chlasta and P. Holas, *Hybrid Approach to Detecting Symptoms of Depression in Social Media Entries*,
Proceedings of the Pacific Asia Conference on Information Systems (2021) (cit. on p. 14).

[134] A. Dinu and A.-C. Moldovan, *Automatic Detection and Classification of Mental Illnesses From General Social Media Texts*,
Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP) (2021) pp. 358–366 (cit. on p. 14).

[135] S. Ji et al., *MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare*, Proceedings of the Thirteenth Language Resources and Evaluation Conference (2022) pp. 7184–7190 (cit. on pp. 14, 66).

[136] S. Ji et al., *Domain-Specific Continued Pretraining of Language Models for Capturing Long Context in Mental Health*,
arXiv preprint arXiv:2304.10447 (2023) (cit. on p. 15).

[137] U. Naseem, B. C. Lee, M. Khushi, J. Kim and A. Dunn,
*Benchmarking for Public Health Surveillance Tasks on Social Media With a Domain-Specific Pretrained Language Model*,
Proceedings of NLP Power! Workshop (2022) pp. 22–31 (cit. on p. 15).

[138] M. Aragon, A. P. L. Monroy, L. Gonzalez, D. E. Losada and M. Montes, *DisorBERT: A Double Domain Adaptation Model for Detecting Signs of Mental Disorders in Social Media*, Proceedings of the Annual Meeting of the Association for Computational Linguistics (2023) pp. 15305–15318 (cit. on p. 15).

[139] E. B. Souto, A. Pérez and J. Parapar,
*Explainability, Interpretability, Depression Detection, Social Media*,
arXiv preprint arXiv:2310.13664 (2023) (cit. on p. 15).

[140] K. Yang, T. Zhang, Z. Kuang, Q. Xie and S. Ananiadou,
*Mentalllama: Interpretable Mental Health Analysis on Social Media With Large Language Models*, arXiv preprint arXiv:2309.13567 (2023)
(cit. on p. 15).

[141] R. Taori et al., *Stanford Alpaca: An Instruction-Following LLaMA Model*,
https://github.com/tatsu-lab/stanford_alpaca, 2023
(cit. on pp. 15, 66).

[142] H. W. Chung et al., *Scaling Instruction-Finetuned Language Models*, Journal of Machine Learning Research 25 (70 2024) pp. 1–53 (cit. on p. 15).

[143] OpenAI, *Introducing ChatGPT*, 2022, URL: https://openai.com/index/chatgpt/ (cit. on p. 15).

[144] S. Bubeck et al., *Sparks of Artificial General Intelligence: Early Experiments With GPT-4*, arXiv preprint arXiv:2303.12712 (2023) (cit. on p. 15).

[145] W.-L. Chiang et al., *Vicuna: An Open-Source Chatbot Impressing GPT-4 With 90%* Chatgpt Quality*, 2023, URL: https://vicuna.lmsys.org/ (cit. on p. 15).

[146] H. Dai et al., *AugGPT: Leveraging ChatGPT for Text Data Augmentation*, arXiv preprint arXiv:2302.13007 (2023) (cit. on pp. 15, 65, 66).

[147] Y.-J. Lee, C.-G. Lim, Y. Choi, J.-H. Lm and H.-J. Choi, *PERSONACHATGEN: Generating Personalized Dialogues Using GPT-3*, Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge (2022) pp. 29–48 (cit. on p. 15).

[148] S. Ubani, S. O. Polat and R. Nielsen, *ZeroShotDataAug: Generating and Augmenting Training Data With ChatGPT*, arXiv preprint arXiv:2304.14334 (2023) (cit. on pp. 15, 66).

[149] S. Meyer, D. Elsweiler, B. Ludwig, M. Fernandez-Pichel and D. E. Losada, *Do We Still Need Human Assessors? Prompt-Based GPT-3 User Simulation in Conversational AI*, Proceedings of the 4th Conference on Conversational User Interfaces (2022) pp. 1–6 (cit. on pp. 15, 65).

[150] T. Brown et al., *Language Models Are Few-Shot Learners*, Advances in Neural Information Processing Systems 33 (2020) pp. 1877–1901 (cit. on pp. 15, 66, 79).

[151] Z. Zhang, S. Chen, M. Wu and K. Zhu, *Symptom Identification for Interpretable Detection of Multiple Mental Disorders on Social Media*, Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (2022) pp. 9970–9985 (cit. on pp. 16, 22).

[152] T. Nguyen, A. Yates, A. Zirikly, B. Desmet and A. Cohan, *Improving the Generalizability of Depression Detection by Leveraging Clinical Questionnaires*, Proceedings of the Annual Meeting of the Association for Computational Linguistics (2022) pp. 8446–8459 (cit. on p. 16).

[153] A. Pérez, J. Parapar, á. Barreiro and S. Lopez-Larrosa, *BDI-Sen: A Sentence Dataset for Clinical Symptoms of Depression*, Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (2023) pp. 2996–3006 (cit. on pp. 16, 23).

[154] A.-S. Uban, B. Chulvi and P. Rosso, *An Emotion and Cognitive Based Analysis of Mental Health Disorders From Social Media Data*, Future Generation Computer Systems (2021) (cit. on pp. 16, 58, 102).

[155] T. Ghosh et al., *An Attention-Based Hybrid Architecture With Explainability for Depressive Social Media Text Detection in Bangla*, Expert Systems with Applications 213 (2023) pp. 119007 (cit. on p. 16).

[156] S. M. Lundberg and S.-I. Lee, *A Unified Approach to Interpreting Model Predictions*, Advances in Neural Information Processing Systems 30 (2017), ed. by I. Guyon et al. (cit. on pp. 16, 39, 60).

[157] M. T. Ribeiro, S. Singh and C. Guestrin, *" Why Should I Trust You?" Explaining the Predictions of Any Classifier*, Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (2016) pp. 1135–1144 (cit. on pp. 16, 60).

[158] V. Adarsh, P. Arun Kumar, V. Lavanya and G. Gangadharan, *Fair and Explainable Depression Detection in Social Media*, Information Processing & Management 60 (2023) pp. 103168 (cit. on pp. 16, 23).

[159] Z. Zhang, S. Chen, M. Wu and K. Zhu, *Psychiatric Scale Guided Risky Post Screening for Early Detection of Depression*, Proceedings of the International Joint Conference on Artificial Intelligence (2022) (cit. on p. 16).

[160] T. Liu et al., *Detecting Symptoms of Depression on Reddit*, Proceedings of ACM Web Science Conference, WebSci '23 (2023) pp. 174–183 (cit. on pp. 16, 23).

[161] G. Rao, Y. Zhang, L. Zhang, Q. Cong and Z. Feng, *MGL-CNN: A Hierarchical Posts Representations Model for Identifying Depressed Individuals in Online Forums*, IEEE Access 8 (2020) pp. 32395–32403 (cit. on p. 16).

[162] M. Trotzek, S. Koitka and C. M. Friedrich, *Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences*, IEEE Transactions on Knowledge and Data Engineering 32 (2018) pp. 588–601 (cit. on pp. 16, 53).

[163] T. Gui et al.,
*Cooperative Multimodal Approach to Depression Detection in Twitter*,
Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) pp. 110–117 (cit. on pp. 17, 22, 74, 76–78, 80, 81, 84, 85, 97, 126).

[164] M. An, J. Wang, S. Li and G. Zhou,
*Multimodal Topic-Enriched Auxiliary Learning for Depression Detection*,
Proceedings of the 28th International Conference on Computational Linguistics (2020) pp. 1078–1089 (cit. on pp. 17, 35, 76, 77, 80, 81).

[165] L. Lustberg and C. F. Reynolds,
*Depression and Insomnia: Questions of Cause and Effect*,
Sleep Medicine Reviews 4 (2000) pp. 253–262 (cit. on p. 17).

[166] M. Stankevich, V. Isakov, D. Devyatkin and I. V. Smirnov,
*Feature Engineering for Depression Detection in Social Media*,
Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (2018) pp. 426–431 (cit. on p. 17).

[167] J. C. Cheng and A. L. Chen,
*Multimodal Time-Aware Attention Networks for Depression Detection*,
Journal of Intelligent Information Systems (2022) pp. 1–21
(cit. on pp. 17, 74, 75, 79–81).

[168] R. Sawhney, H. Joshi, S. Gandhi and R. Shah, *A Time-Aware Transformer Based Model for Suicide Ideation Detection on Social Media*,
"Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (2020) pp. 7685–7697 (cit. on pp. 17, 75, 79).

[169] A. Pampouchidou et al., *Automatic Assessment of Depression Based on Visual Cues: A Systematic Review*,
IEEE Transactions on Affective Computing 10 (2017) pp. 445–470
(cit. on p. 17).

[170] E. Friesen and P. Ekman, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*,
Palo Alto University, California 3 (1978) pp. 5 (cit. on pp. 17, 90).

[171] A. Pampouchidou et al., *Depression Assessment by Fusing High and Low Level Features From Audio, Video, and Text*,
Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (2016) pp. 27–34 (cit. on p. 17).

[172] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli and D. D. Mehta, *Vocal and Facial Biomarkers of Depression Based on Motor Incoordination and Timing*, Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (2014) pp. 65–72 (cit. on p. 17).

[173] Y. Tao et al., *Depressive Semantic Awareness From Vlog Facial and Vocal Streams via Spatio-Temporal Transformer*, Digital Communications and Networks (2023) (cit. on pp. 17, 92, 93, 95, 97, 127).

[174] L. Zhou et al., *CAIINET: Neural Network Based on Contextual Attention and Information Interaction Mechanism for Depression Detection*, Digital Signal Processing 137 (2023) pp. 103986 (cit. on pp. 17, 85, 88, 92, 93, 97, 127).

[175] L. Zhou et al., *TAMFN: Time-Aware Attention Multimodal Fusion Network for Depression Detection*, IEEE Transactions on Neural Systems and Rehabilitation Engineering 31 (2022) pp. 669–679 (cit. on pp. 17, 88, 92, 93, 97, 127).

[176] W. Zheng, L. Yan and F.-Y. Wang, *Two Birds With One Stone: Knowledge-Embedded Temporal Convolutional Transformer for Depression Detection and Emotion Recognition*, IEEE Transactions on Affective Computing (2023) pp. 1–18 (cit. on pp. 17, 88, 92, 93, 95–97, 127).

[177] J. Yoon, C. Kang, S. Kim and J. Han, *D-Vlog: Multimodal Vlog Dataset for Depression Detection*, Proceedings of the AAAI Conference on Artificial Intelligence 36 (2022) pp. 12226–12234 (cit. on pp. 17, 35, 85, 87, 88, 90, 92, 93, 97, 127).

[178] D.-K. Nguyen et al., *Multimodal Transformer for Automatic Depression Estimation System*, Proceedings of the 29th International Workshop on Frontiers of Computer Vision (2023) (cit. on pp. 17, 85, 88, 92, 93, 97, 127).

[179] D. Zhu, J. Chen, X. Shen, X. Li and M. Elhoseiny, *Minigpt-4: Enhancing Vision-Language Understanding With Advanced Large Language Models*, Proceedings of the Twelfth International Conference on Learning Representations (2023) (cit. on p. 17).

[180] Y. Su et al., *PandaGPT: One Model to Instruction-Follow Them All*, Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants (2023) pp. 11–23 (cit. on p. 17).

[181] D. E. Losada, F. A. Crestani and J. Parapar, *Overview of eRisk 2018: Early Risk Prediction on the Internet (Extended Lab Overview)*, Working Notes of CLEF 2018, Early Risk Prediction on the Internet Lab (eRisk), CEUR 2125 (2018) (cit. on pp. 17, 21, 33, 38, 46).

[182] D. E. Losada, F. Crestani and J. Parapar, *Overview of Erisk 2019 Early Risk Prediction on the Internet*, Working Notes of CLEF 2019, Early Risk Prediction on the Internet Lab (eRisk) (2019) pp. 340–357 (cit. on pp. 17, 22, 74).

[183] D. E. Losada, F. Crestani and J. Parapar,
*Overview of ERisk 2020: Early Risk Prediction on the Internet*,
Experimental IR Meets Multilinguality, Multimodality, and Interaction:
11th International Conference of the CLEF Association, CLEF 2020,
Thessaloniki, Greece, September 22–25, 2020, Proceedings (2020)
pp. 272–287 (cit. on pp. 17, 22, 53).

[184] F. Crestani, D. E. Losada and J. Parapar,
*Early Detection of Mental Health Disorders by Social Media Monitoring:
The First Five Years of the eRisk Project*, vol. 1018, Springer Nature, 2022
(cit. on pp. 17, 23).

[185] K. Harrigian, C. Aguirre and M. Dredze,
*On the State of Social Media Data for Mental Health Research*,
Proceedings of the Workshop on Computational Linguistics and Clinical
Psychology (CLPsych), NAACL (2021) pp. 15–24 (cit. on p. 18).

[186] M. M. Aldarwish and H. F. Ahmad,
*Predicting Depression Levels Using Social Media Posts*,
2017 IEEE 13th international Symposium on Autonomous decentralized
system (ISADS) (2017) pp. 277–280 (cit. on p. 21).

[187] D. Mowery et al., *Understanding Depressive Symptoms and Psychosocial
Stressors on Twitter: A Corpus-Based Study*,
Journal of Medical Internet Research 19 (2017) e6895 (cit. on p. 21).

[188] A. H. Yazdavar et al., *Semi-Supervised Approach to Monitoring Clinical
Depressive Symptoms in Social Media*,
Proceedings of the IEEE/ACM International Conference on Advances in
Social Network Analysis and Mining. International Conference on
Advances in Social Network Analysis and Mining (2017) pp. 1191–1198
(cit. on p. 21).

[189] L. Rojas-Barahona et al., *Deep Learning for Language Understanding of
Mental Health Concepts Derived From Cognitive Behavioural Therapy*,
Proceedings of the Ninth International Workshop on Health Text Mining
and Information Analysis (2018) pp. 44–54 (cit. on p. 21).

[190] E. M. Seabrook, M. L. Kern, B. D. Fulcher and N. S. Rickard,
*Predicting Depression From Language-Based Emotion Dynamics:
Longitudinal Analysis of Facebook and Twitter Status Updates*,
Journal of Medical Internet Research 20 (2018) e9267 (cit. on p. 21).

[191] B. J. Ricard, L. A. Marsch, B. Crosier and S. Hassanpour,
*Exploring the Utility of Community-Generated Social Media Content for
Detecting Depression: An Analytical Study on Instagram*,
Journal of Medical Internet Research 20 (2018) e11817 (cit. on p. 21).

[192] X. Chen, M. D. Sykora, T. W. Jackson, S. Elayan and F. Munir, *Tweeting Your Mental Health: An Exploration of Different Classifiers and Features With Emotional Signals in Identifying Mental Health Conditions*, Hawaii International Conference on System Sciences (2018) (cit. on p. 21).

[193] X. Chen, M. D. Sykora, T. W. Jackson and S. Elayan, *What About Mood Swings: Identifying Depression on Twitter With Temporal Measures of Emotions*, Companion Proceedings of the The Web Conference 2018, WWW '18 (2018) pp. 1653–1660 (cit. on p. 21).

[194] S. MacAvaney et al., *Rsdd-Time: Temporal Annotation of Self-Reported Mental Health Diagnoses*, Proceedings of the Workshop on Computational Linguistics and Clinical Psychology (CLPsych), NAACL (2018) pp. 168–173 (cit. on p. 21).

[195] M. Islam et al., *Depression Detection From Social Network Data Using Machine Learning Techniques*, Health Information Science and Systems 6 (2018) pp. 1–12 (cit. on p. 21).

[196] S. Almouzini, A. Alageel et al., *Detecting Arabic Depressed Users From Twitter Data*, Procedia Computer Science 163 (2019) pp. 257–265 (cit. on p. 22).

[197] K. C. Bathina, M. Ten Thij, L. Lorenzo-Luaces, L. A. Rutter and J. Bollen, *Individuals With Depression Express More Distorted Thinking on Social Media*, Nature Human Behaviour 5 (2021) pp. 458–466 (cit. on p. 22).

[198] E. A. Ríssola, S. A. Bahrainian and F. Crestani, *A Dataset for Research on Depression in Social Media*, Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (2020) (cit. on pp. 22, 46, 47, 57, 61, 64, 84).

[199] M. Birnbaum et al., *Identifying Signals Associated With Psychiatric Illness Utilizing Language and Images Posted to Facebook*, npj Schizophrenia 6 (2020) (cit. on p. 22).

[200] S. Yadav et al., *Identifying Depressive Symptoms From Tweets: Figurative Language Enabled Multitask Learning Framework*, Proceedings of the 28th International Conference on Computational Linguistics (2020) pp. 696–709 (cit. on p. 22).

[201] T. Tabak and M. Purver, *Temporal Mental Health Dynamics on Social Media*, Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020 (2020) (cit. on pp. 22, 26, 28, 29).

[202] E. Sherman, K. Harrigian, C. Aguirre and M. Dredze, *Towards Understanding the Role of Gender in Deploying Social Media-Based Mental Health Surveillance Models*, Proceedings of the Workshop on Computational Linguistics and Clinical Psychology (CLPsych), NAACL (2021) pp. 217–223 (cit. on p. 22).

[203] J. Pirayesh, H. Chen, X. Qin, W.-S. Ku and D. Yan, *MentalSpot: Effective Early Screening for Depression Based on Social Contagion*, Proceedings of the 30th ACM International Conference on Information & Knowledge Management, CIKM '21 (2021) pp. 1437–1446 (cit. on p. 22).

[204] X. Guo, Y. Sun and S. Vosoughi, *Emotion-Based Modeling of Mental Disorders on Social Media*, Proceedings of IEEE/WIC/ACM WI-IAT (2021) pp. 8–16 (cit. on p. 22).

[205] Y. Zhang et al., *Monitoring Depression Trends on Twitter During the COVID-19 Pandemic: Observational Study*, JMIR Infodemiology 1 (2021) e26769 (cit. on pp. 22, 26, 28).

[206] J. Zhou et al., *Detecting Community Depression Dynamics Due to Covid-19 Pandemic in Australia*, IEEE Transactions on Computational Social Systems 8 (2021) pp. 982–991 (cit. on pp. 22, 26, 28, 30).

[207] R. Safa, P. Bayat and L. Moghtader, *Automatic Detection of Depression Symptoms in Twitter Using Multimodal Analysis*, The Journal of Supercomputing 78 (2022) (cit. on p. 22).

[208] S. Boinepelli et al., *Leveraging Mental Health Forums for User-Level Depression Detection on Social Media*, Proceedings of the Thirteenth Language Resources and Evaluation Conference (2022) pp. 5418–5427 (cit. on p. 22).

[209] M. Garg et al., *Cams: An Annotated Corpus for Causal Analysis of Mental Health Issues in Social Media Posts*, Proceedings of the Thirteenth Language Resources and Evaluation Conference (2022) pp. 6387–6396 (cit. on p. 22).

[210] S. Sotudeh, N. Goharian and Z. Young, *MentSum: A Resource for Exploring Summarization of Mental Health Online Posts*, Proceedings of the Thirteenth Language Resources and Evaluation Conference (2022) pp. 2682–2692 (cit. on p. 22).

[211] A. Monreale, B. Iavarone, E. Rossetto and A. Beretta, *Detecting Addiction, Anxiety, and Depression by Users Psychometric Profiles*, Companion Proceedings of the Web Conference 2022, WWW '22 (2022) pp. 1189–1197 (cit. on p. 23).

[212] S. Gupta et al., *Learning to Automate Follow-Up Question Generation Using Process Knowledge for Depression Triage on Reddit Posts*, Proceedings of the Workshop on Computational Linguistics and Clinical Psychology (CLPsych), NAACL (2022) pp. 137–147 (cit. on p. 23).

[213] I. Mihov et al., *MentalNet: Heterogeneous Graph Representation for Early Depression Detection*, Proceedings of the 2022 IEEE International Conference on Data Mining (2022) pp. 1113–1118 (cit. on p. 23).

[214] A. K. Singh et al., *Twitter-Stmhd: An Extensive User-Level Database of Multiple Mental Health Disorders*, Proceedings of the International AAAI Conference on Web and Social Media 16 (2022) pp. 1182–1191 (cit. on p. 23).

[215] B. Davis et al., *Quantifying Depression-Related Language on Social Media During the COVID-19 Pandemic*, International Journal of Population Data Science 5 (2022) (cit. on p. 23).

[216] I. Fernández-Barrera, S. Bravo-Bustos and M. Vidal, *Evaluating the Social Media Users' Mental Health Status During COVID-19 Pandemic Using Deep Learning*, International Conference on Biomedical and Health Informatics 14 (2022) (cit. on pp. 23, 27, 28, 85).

[217] J. Cha, S. Kim and E. Park, *A Lexicon-Based Approach to Examine Depression Detection in Social Media: The Case of Twitter and University Community*, Humanities and Social Sciences Communications 9 (2022) pp. 1–10 (cit. on pp. 23, 32).

[218] S. Z. Alavijeh, F. Zarrinkalam, Z. Noorian, A. Mehrpour and K. Etminani, *What Users' Musical Preference on Twitter Reveals About Psychological Disorders*, Information Processing & Management 60 (2023) pp. 103269 (cit. on p. 23).

[219] H. Song, J. Shin, H. Lee and J. C. Park, *A Simple and Flexible Modeling for Mental Disorder Detection by Learning From Clinical Questionnaires*, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2023) pp. 12190–12206 (cit. on p. 23).

[220] Y. Liang, L. Liu, Y. Ji, L. Huangfu and D. D. Zeng, *Identifying Emotional Causes of Mental Disorders From Social Media for Effective Intervention*, Information Processing & Management 60 (2023) pp. 103407 (cit. on p. 23).

[221] Y. Liu, L. Biester and R. Mihalcea, *Improving Mental Health Classifier Generalization With Pre-Diagnosis Data*, Proceedings of the International AAAI Conference on Web and Social Media 17 (2023) pp. 566–577 (cit. on p. 23).

[222] S. Yadav et al., *Towards Identifying Fine-Grained Depression Symptoms From Memes*, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (2023) pp. 8890–8905 (cit. on pp. 23, 85).

[223] H. Zogan, I. Razzak, S. Jameel and G. Xu,
*Hierarchical Convolutional Attention Network for Depression Detection on Social Media and Its Impact During Pandemic*,
IEEE Journal of Biomedical and Health Informatics (2023)
(cit. on pp. 23, 32).

[224] J. Wu et al., *Exploring Social Media for Early Detection of Depression in COVID-19 Patients*,
Proceedings of ACM Web Conference (2023) pp. 3968–3977
(cit. on pp. 23, 28, 31).

[225] M. E. Villa-Pérez, L. A. Trejo, M. B. Moin and E. Stroulia,
*Extracting Mental Health Indicators From English and Spanish Social Media: A Machine Learning Approach*,
IEEE Access 11 (2023) pp. 128135–128152 (cit. on p. 23).

[226] Y. Priyadarshana, Z. Liang and I. Piumarta,
*HelaDepDet: A Novel Multi-Class Classification Model for Detecting the Severity of Human Depression*, International Conference on Collaboration Technologies and Social Computing (2023) pp. 3–18 (cit. on p. 23).

[227] A. Anshul, G. S. Pranav, M. Z. U. Rehman and N. Kumar,
*A Multimodal Framework for Depression Detection During COVID-19 via Harvesting Social Media*,
IEEE Transactions on Computational Social Systems (2023) (cit. on p. 24).

[228] F. Alhamed, J. Ive and L. Specia,
*Classifying Social Media Users Before and After Depression Diagnosis via Their Language Usage: A Dataset and Study*, Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (2024) pp. 3250–3260
(cit. on p. 24).

[229] K. Milintsevich, K. Sirts and G. Dias, *Your Model Is Not Predicting Depression Well and That Is Why: A Case Study of PRIMATE Dataset*,
Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024) (2024) pp. 166–171 (cit. on p. 24).

[230] D. Lee et al., *Detecting Bipolar Disorder From Misdiagnosed Major Depressive Disorder With Mood-Aware Multi-Task Learning*,
Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (2024) pp. 4954–4970 (cit. on p. 24).

[231] R. Beniwal and P. Saraswat, *A Hybrid BERT-CNN Approach for Depression Detection on Social Media Using Multimodal Data*,
The Computer Journal (2024) bxae018 (cit. on p. 24).

[232] F. B. Tumaliuan, L. Grepo and E. R. Jalao, *Development of Depression Data Sets and a Language Model for Depression Detection: Mixed Methods Study*, JMIR Data 5 (2024) e53365 (cit. on p. 24).

[233] J. Wolohan, *Estimating the Effect of COVID-19 on Mental Health: Linguistic Indicators of Depression During a Global Pandemic*, Proceedings of NLP-COVID19 Workshop, EMNLP (2020) (cit. on pp. 25–28).

[234] A. Kaseb, O. Galal and D. Elreedy, *Analysis on Tweets Towards COVID-19 Pandemic: An Application of Text-Based Depression Detection*, Proceedings of the 4th Novel Intelligent and Leading Emerging Sciences Conference (2022) pp. 131–136 (cit. on pp. 25, 26, 28).

[235] M. Lenharo, *WHO Declares End to COVID-19's Emergency Phase*, Nature 882 (2023) (cit. on p. 25).

[236] J. F. Bankston and L. Ma, *A Study on People's Mental Health on Twitter During the COVID-19 Pandemic*, Proceedings of the 2022 6th International Conference on Compute and Data Analysis (2022) pp. 111–115 (cit. on pp. 26, 28).

[237] A. Tommasel, A. Diaz-Pace, D. Godoy and J. M. Rodriguez, *Tracking the Evolution of Crisis Processes and Mental Health on Social Media During the COVID-19 Pandemic*, Behaviour & Information Technology 41 (2022) pp. 3450–3469 (cit. on pp. 26, 28, 29).

[238] L. Biester, K. Matton, J. Rajendran, E. M. Provost and R. Mihalcea, *Quantifying the Effects of COVID-19 on Mental Health Support Forums*, Proceedings of NLP-COVID19 Workshop, EMNLP (2020) (cit. on pp. 26–28).

[239] A. Fine, P. Crutchley, J. Blase, J. Carroll and G. Coppersmith, *Assessing Population-Level Symptoms of Anxiety, Depression, and Suicide Risk in Real Time Using NLP Applied to Social Media Data*, Proceedings of NLP+CSS Workshop (2020) pp. 50–54 (cit. on pp. 26, 28, 31).

[240] I. Li et al., *What Are We Depressed About When We Talk About COVID-19: Mental Health Analysis on Tweets Using Natural Language Processing*, International Conference on Innovative Techniques and Applications of Artificial Intelligence (2020) pp. 358–370 (cit. on pp. 26, 28).

[241] L. Biester, K. Matton, J. Rajendran, E. M. Provost and R. Mihalcea, *Understanding the Impact of COVID-19 on Online Mental Health Forums*, ACM Transactions on Management Information Systems 12 (2021) pp. 1–28 (cit. on pp. 27, 28).

[242] J. Q. Chen, K. Qi, A. Zhang, M. Shalaginov and T. H. Zeng, *COVID-19 Impact on Mental Health Analysis Based on Reddit Comments*, Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (2022) pp. 2253–2258 (cit. on p. 27).

[243] G. S. Bajaj, H. Yadav, H. S. Sahdev, S. Sah and P. Kaur, *Mental Health Analysis During COVID-19: A Comparison Before and During the Pandemic*, Proceedings of the IEEE 4th International Conference on Computing, Power and Communication Technologies (2021) pp. 1–7 (cit. on pp. 27, 28).

[244] P. C. Sukhwal and A. Kankanhalli, *Determining Containment Policy Impacts on Public Sentiment During the Pandemic Using Social Media Data*, Proceedings of the National Academy of Sciences 119 (2022) e2117292119 (cit. on pp. 27, 28, 30).

[245] J. Billings, B. C. F. Ching, V. Gkofa, T. Greene and M. Bloomfield, *Experiences of Frontline Healthcare Workers and Their Views About Support During COVID-19 and Previous Pandemics: A Systematic Review and Qualitative Meta-Synthesis*, BMC health services research 21 (2021) pp. 1–17 (cit. on p. 28).

[246] D. Li, H. Chaudhary and Z. Zhang, *Modeling Spatiotemporal Pattern of Depressive Symptoms Caused by COVID-19 Using Social Media Data Mining*, International Journal of Environmental Research and Public Health 17 (2020) pp. 4988 (cit. on p. 28).

[247] K. Saha, J. Torous, E. D. Caine and M. De Choudhury, *Psychosocial Effects of the COVID-19 Pandemic: Large-Scale Quasi-Experimental Study on Social Media*, Journal of Medical Internet Research 22 (2020) e22600 (cit. on pp. 28, 29).

[248] S. Zhang et al., *The COVID-19 Pandemic and Mental Health Concerns on Twitter in the United States*, Health Data Science (2022) (cit. on p. 28).

[249] M. Li et al., *Tracking the Impact of COVID-19 and Lockdown Policies on Public Mental Health Using Social Media: Infoveillance Study*, Journal of Medical Internet Research 24 (2022) e39676 (cit. on pp. 28, 31).

[250] D. Levanti et al., *Depression and Anxiety on Twitter During the COVID-19 Stay-at-Home Period in 7 Major US Cities*, AJPM focus 2 (2023) pp. 100062 (cit. on p. 28).

[251] J. M. Tshimula, B. Chikhaoui and S. Wang, *COVID-19: Detecting Depression Signals During Stay-at-Home Period*, Health Informatics Journal 28 (2022) pp. 14604582221094931 (cit. on pp. 28, 29).

[252] C. Marshall et al.,
*Using Natural Language Processing to Explore Mental Health Insights From UK Tweets During the COVID-19 Pandemic: Infodemiology Study*,
Jmir Infodemiology 2 (2022) e32449 (cit. on pp. 28, 29).

[253] M. K. Bashar, *Tracking Public Depression From Tweets on COVID-19 and Its Comparison With Pre-Pandemic Time*,
Proceedings of the International Conference on Intelligent Informatics and Biomedical Science 7 (2022) pp. 389–395 (cit. on pp. 28, 30).

[254] G. Gour, V. S. Savantanavar, Yashoda, V. Gadyal and S. Basavaraddi,
*Depression Analysis of Real Time Tweets During Covid Pandemic*,
Proceedings of the International Conference on Ubiquitous Computing and Intelligent Information Systems (2021) pp. 55–73 (cit. on pp. 28, 30).

[255] B. D. Davis et al., *Quantifying Depression-Related Language on Social Media During the COVID-19 Pandemic*,
International Journal of Population Data Science 5 (2020) (cit. on p. 28).

[256] O. El-Gayar, A. Wahbeh, T. Nasralah, A. Elnoshokaty and A.-R. Mohammad,
*Mental Health and the COVID-19 Pandemic: Analysis of Twitter Discourse*,
omputer Information Systems Faculty Publications 4 (2021) (cit. on p. 28).

[257] S. Nandy and V. Kumar, *My Mind Is a Prison: A Boosted Deep Learning Approach to Detect the Rise in Depression Since COVID-19 Using a Stacked Bi-LSTM CatBoost Model*,
Proceedings of Big Data (2021) pp. 4396–4400 (cit. on p. 28).

[258] U. M. Bello et al., *Prevalence of Anxiety and Depression Among the General Population in Africa During the COVID-19 Pandemic: A Systematic Review and Meta-Analysis*,
Frontiers in Public Health 10 (2022) pp. 814981 (cit. on p. 28).

[259] H. Wang, L. Ma and Y. Yu, *Monitoring Worldwide Trends of Expressed Depression on Twitter Before and After COVID-19 Vaccine Releases*,
Proceedings of the 10th International Conference on Bioinformatics and Computational Biology (2022) pp. 148–153 (cit. on pp. 28, 30).

[260] J. Eisenstein, A. Ahmed and E. P. Xing,
*Sparse Additive Generative Models of Text*, Proceedings of the International Conference on Machine Learning (2011) pp. 1041–1048 (cit. on p. 29).

[261] M. S. Spoorthy, S. K. Pratapa and S. Mahant, *Mental Health Problems Faced by Healthcare Workers Due to the COVID-19 Pandemic–a Review*,
Asian Journal of Psychiatry 51 (2020) pp. 102119 (cit. on p. 31).

[262] A. K. Agarwal, J. Mittal, A. Tran, R. Merchant and S. C. Guntuku, *Investigating Social Media to Evaluate Emergency Medicine Physicians' Emotional Well-Being During COVID-19*, JAMA Network Open 6 (2023) e2312708–e2312708 (cit. on p. 31).

[263] B. Anbalagan, *Detecting Depressive Online User Behavior During Global Pandemic by Fusing LSTM and CNN Models*, Proceedings of ICAIAA (2022) pp. 1–10 (cit. on p. 32).

[264] R. Meena and V. T. Bai, *Depression Detection on COVID 19 Tweets Using Chimp Optimization Algorithm.*, Intelligent Automation & Soft Computing 34 (2022) (cit. on p. 32).

[265] M. H. Al Banna et al., *A Hybrid Deep Learning Model to Predict the Impact of COVID-19 on Mental Health From Social Media Big Data*, IEEE Access (2023) (cit. on p. 32).

[266] K. Harrigian and M. Dredze, *The Problem of Semantic Shift in Longitudinal Monitoring of Social Media: A Case Study on Mental Health During the COVID-19 Pandemic*, Proceedings of ACM Web Science Conference (2022) pp. 208–218 (cit. on pp. 32, 33).

[267] J. Parapar, P. Martín-Rodilla, D. E. Losada and F. Crestani, "Overview of eRisk 2022: Early Risk Prediction On the Internet", ed. by A. Barrón-Cedeño et al., Cham: Springer International Publishing, 2022 pp. 233–256 (cit. on pp. 34, 104, 105).

[268] J. Gratch et al., *The Distress Analysis Interview Corpus of Human and Computer Interviews*, Proceedings of the International Conference on Language Resources and Evaluation (2014) pp. 3123–3128 (cit. on pp. 35, 36, 85, 90, 97, 127, 128).

[269] K. Kroenke et al., *The PHQ-8 as a Measure of Current Depression in the General Population*, Journal of affective disorders 114 (2009) pp. 163–173 (cit. on p. 36).

[270] F. Ringeval et al., *AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression With AI, and Cross-Cultural Affect Recognition*, Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (2019) pp. 3–12 (cit. on pp. 36, 85, 90, 97, 127, 128).

[271] A. Taylor, M. Marcus and B. Santorini, *The Penn Treebank: An Overview*, Treebanks (2003) pp. 5–22 (cit. on p. 38).

[272] C. Caragea, L. P. Dinu and B. Dumitru, *Exploring Optimism and Pessimism in Twitter Using Deep Learning*, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (2018) pp. 652–658 (cit. on p. 38).

[273] J. M. Havigerová, J. Haviger, D. Kučera and P. Hoffmannová, *Text-Based Detection of the Risk of Depression*, Frontiers in Psychology 10 (2019) pp. 513 (cit. on p. 39).

[274] T. Litvinova, O. Zagorovskaya, O. Litvinova and P. Seredin, *Profiling a Set of Personality Traits of a Text's Author: A Corpus-Based Approach*, International Conference on Speech and Computer (2016) pp. 555–562 (cit. on pp. 39, 42).

[275] F. Mairesse, M. A. Walker, M. R. Mehl and R. K. Moore, *Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text*, Journal of Artificial Intelligence Research 30 (2007) pp. 457–500 (cit. on p. 39).

[276] A. Kilgarriff, *Simple Maths for Keywords*, Proceedings of the Corpus Linguistics Conference (2009) (cit. on pp. 39, 48).

[277] C. Gabrielatos, *Keyness Analysis*, Routledge London (2018) pp. 225–258 (cit. on pp. 39, 48).

[278] T. E. Dunning, *Accurate Methods for the Statistics of Surprise and Coincidence*, Computational linguistics 19 (1993) pp. 61–74 (cit. on pp. 40, 48).

[279] S. Nolen-Hoeksema, *Responses to Depression and Their Effects on the Duration of Depressive Episodes.*, Journal of Abnormal Psychology 100 (1991) pp. 569 (cit. on p. 40).

[280] M. De Choudhury and S. De, *Mental Health Discourse on Reddit: Self-Disclosure, Social Support, and Anonymity*, Proceedings of the International AAAI Conference on Web and Social Media 8 (2014) (cit. on pp. 41, 42, 66, 77, 83, 122).

[281] D. R. Hopko, M. E. Armento, M. S. Cantu, L. L. Chambers and C. Lejuez, *The Use of Daily Diaries to Assess the Relations Among Mood State, Overt Behavior, and Reward Value of Activities*, Behaviour Research and Therapy 41 (2003) pp. 1137–1148 (cit. on p. 41).

[282] J. Pennebaker, *Mind Mapping: Using Everyday Language to Explore Social & Psychological Processes*, Procedia Computer Science 118 (2017) pp. 100–107 (cit. on p. 42).

[283] T. Brockmeyer et al., *Me, Myself, and I: Self-Referent Word Use as an Indicator of Self-Focused Attention in Relation to Depression and Anxiety*, Frontiers in Psychology 6 (2015) pp. 1564 (cit. on pp. 43, 48).

[284] D. Premack and G. Woodruff, *Does the Chimpanzee Have a Theory of Mind?*, Behavioral and Brain Sciences 1 (1978) pp. 515–526 (cit. on p. 43).

[285] T. M. Erle, N. Barth and S. Topolinski,
*Egocentrism in Sub-Clinical Depression*,
Cognition and Emotion 33 (2019) pp. 1239–1248 (cit. on p. 43).

[286] M. F. Steger and T. B. Kashdan,
*Depression and Everyday Social Activity, Belonging, and Well-Being.*,
Journal of Counseling Psychology 56 (2009) pp. 289 (cit. on p. 43).

[287] F. Heylighen and J.-M. Dewaele,
*Variation in the Contextuality of Language: An Empirical Measure*,
Foundations of Science 7 (2002) pp. 293–340 (cit. on p. 44).

[288] Z. Waseem, T. Davidson, D. Warmsley and I. Weber,
*Understanding Abuse: A Typology of Abusive Language Detection Subtasks*,
Proceedings of the First Workshop on Abusive Language Online (2017)
(cit. on p. 45).

[289] V. Basile et al., *Semeval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter*, Proceedings of the 13th International Workshop on Semantic Evaluation (2019) (cit. on p. 45).

[290] S. Malmasi and M. Zampieri,
*Challenges in Discriminating Profanity From Hate Speech*, Journal of Experimental & Theoretical Artificial Intelligence 30 (2 2018) pp. 1–16
(cit. on p. 45).

[291] E. Holgate, I. Cachola, D. Preotiuc-Pietro and J. J. Li,
*Why Swear? Analyzing and Inferring the Intentions of Vulgar Expressions*,
Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (2018) (cit. on p. 45).

[292] M. Zampieri et al.,
*Predicting the Type and Target of Offensive Posts in Social Media*,
Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (2019) (cit. on pp. 45, 46).

[293] M. Zampieri et al., *SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)*,
Proceedings of the 13th International Workshop on Semantic Evaluation (2019) pp. 75–86, ed. by J. May et al. (cit. on p. 45).

[294] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*,
Proceedings of the International Conference on Learning Representations (2015) (cit. on pp. 46, 60, 80, 107, 111).

[295] S. Rosenthal, P. Atanasova, G. Karadzhov, M. Zampieri and P. Nakov,
*A Large-Scale Semi-Supervised Dataset for Offensive Language Identification*, Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (2021) (cit. on p. 46).

[296] S. M. Mohammad and P. D. Turney, *Crowdsourcing a Word–Emotion Association Lexicon*, Computational Intelligence 29 (2013) pp. 436–465 (cit. on p. 46).

[297] S. M. Mohammad, *Word Affect Intensities*, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (2017) (cit. on p. 46).

[298] J. J. Gross, *Emotion and Emotion Regulation*, Handbook of Personality: Theory and Research 2 (1999) pp. 525–552 (cit. on p. 48).

[299] R. Stephens and O. Robertson, *Swearing as a Response to Pain: Assessing Hypoalgesic Effects of Novel "Swear" Words*, Frontiers in Psychology 11 (2020) pp. 723 (cit. on pp. 48, 50).

[300] J. Everaert, I. R. Podina and E. H. Koster, *A Comprehensive Meta-Analysis of Interpretation Biases in Depression*, Clinical Psychology Review 58 (2017) pp. 33–48 (cit. on pp. 48, 50).

[301] S. A. Speer, *Reconsidering Self-Deprecation as a Communication Practice*, British Journal of Social Psychology 58 (2019) pp. 806–828 (cit. on p. 48).

[302] A. T. Beck and E. A. Haigh, *Advances in Cognitive Theory and Therapy: The Generic Cognitive Model*, Annual Review of Clinical Psychology 10 (2014) pp. 1–24 (cit. on pp. 48–50).

[303] J. Manohar, T. Sathyanarayana Rao, S. Chandran et al., *Sexual Dysfunctions in Depression*, Clin Depress 3 (2017) pp. 1–5 (cit. on p. 49).

[304] N. Adami et al., *Markers of Unsustainable Gambling for Early Detection of at-Risk Online Gamblers*, International Gambling Studies 13 (2013) pp. 188–204 (cit. on p. 52).

[305] F. Peres et al., *Time Series Clustering of Online Gambling Activities for Addicted Users' Detection*, Applied Sciences 11 (2021) pp. 2397 (cit. on p. 52).

[306] R. Ladouceur, P. Shaffer, A. Blaszczynski and H. J. Shaffer, *Responsible Gambling: A Synthesis of the Empirical Evidence*, Addiction Research & Theory 25 (2017) pp. 225–235 (cit. on p. 52).

[307] J. Haefeli, S. Lischer and J. Schwarz, *Early Detection Items and Responsible Gambling Features for Online Gambling*, International Gambling Studies 11 (2011) pp. 273–288 (cit. on p. 52).

[308] H. Fliege, J.-R. Lee, A. Grimm and B. F. Klapp, *Risk Factors and Correlates of Deliberate Self-Harm Behavior: A Systematic Review*, Journal of psychosomatic research 66 (2009) pp. 477–493 (cit. on p. 52).

[309] R. Martínez-Castaño, A. Htait, L. Azzopardi and Y. Moshfeghi, *Early Risk Detection of Self-Harm and Depression Severity Using BERT-based Transformers: iLab at CLEF eRisk 2020*, Working Notes of CLEF 2020, Early Risk Prediction on the Internet Lab (eRisk), CEUR 2696 (2020) (cit. on pp. 52, 53).

[310] A. Conneau et al., *Unsupervised Cross-Lingual Representation Learning at Scale*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8440–8451 (2019) (cit. on pp. 52, 79).

[311] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire and J. Blackburn, *The Pushshift Reddit Dataset*, Proceedings of the International AAAI Conference on Web and Social Media 14 (2020) pp. 830–839 (cit. on p. 52).

[312] A. Trifan, P. Salgado and L. Oliveira, *BioInfo@ UAVR at eRisk 2020: On the Use of Psycholinguistics Features and Machine Learning for the Classification and Quantification of Mental Diseases*, Working Notes of CLEF 2020, Early Risk Prediction on the Internet Lab (eRisk), CEUR 2696 (2020) (cit. on p. 52).

[313] L. Achilles, M. Kisselew, J. Schäfer and R. Koelle, *Using Surface and Semantic Features for Detecting Early Signs of Self-Harm in Social Media Postings*, Working Notes of CLEF 2020, Early Risk Prediction on the Internet Lab (eRisk), CEUR 2696 (2020) (cit. on p. 52).

[314] M. E. Aragón, A. P. López-Monroy and M. Montes-y-Gómez, *INAOE-CIMAT at eRisk 2020: Detecting Signs of Self-Harm Using Sub-Emotions and Words*, Working Notes of CLEF 2020, Early Risk Prediction on the Internet Lab (eRisk), CEUR 2696 (2020) (cit. on p. 52).

[315] M. M. Tadesse, H. Lin, B. Xu and L. Yang, *Detection of Depression-Related Posts in Reddit Social Media Forum*, IEEE Access 7 (2019) pp. 44883–44893 (cit. on p. 57).

[316] M. Sundararajan, A. Taly and Q. Yan, *Axiomatic Attribution for Deep Networks*, International Conference on Machine Learning (2017) pp. 3319–3328 (cit. on pp. 57, 60, 61, 83, 84, 96).

[317] D. Maupomé, M. D. Armstrong, F. Rancourt, T. Soulas and M.-J. Meurs, *Early Detection of Signs of Pathological Gambling, Self-Harm and Depression Through Topic Extraction and Neural Networks.*, Working Notes of CLEF 2021, Early Risk Prediction on the Internet Lab (eRisk), CEUR 2936 (2021) pp. 1031–1045 (cit. on pp. 57, 101).

[318] J. M. Loyola, S. Burdisso, H. Thompson, L. Cagnina and M. Errecalde, *UNSL at eRisk 2021: A Comparison of Three Early Alert Policies for Early Risk Detection*, Working Notes of CLEF 2021, Early Risk Prediction on the Internet Lab (eRisk), CEUR 2936 (2021) pp. 992–1021 (cit. on p. 57).

[319] M. Trotzek, S. Koitka and C. M. Friedrich, *Linguistic Metadata Augmented Classifiers at the CLEF 2017 Task for Early Detection of Depression*, Working Notes of CLEF 2017, Early Risk Prediction on the Internet Lab (eRisk), CEUR 1866 (2017) (cit. on p. 58).

[320] M. L. Errecalde, M. P. Villegas, D. G. Funez, M. J. G. Ucelay and L. C. Cagnina, *Temporal Variation of Terms as Concept Space for Early Risk Prediction*, Working Notes of CLEF 2017, Early Risk Prediction on the Internet Lab (eRisk), CEUR 1866 (2017) (cit. on p. 58).

[321] D. G. Funez et al., *UNSL's Participation at eRisk 2018 Lab.*, Working Notes of CLEF 2018, Early Risk Prediction on the Internet Lab (eRisk), CEUR 2125 (2018) (cit. on p. 58).

[322] M. Trotzek, S. Koitka and C. M. Friedrich, *Word Embeddings and Linguistic Metadata at the CLEF 2018 Tasks for Early Detection of Depression and Anorexia.*, Working Notes of CLEF 2018, Early Risk Prediction on the Internet Lab (eRisk), CEUR 2125 (2018) (cit. on p. 58).

[323] M. E. Aragon, A. P. Lopez-Monroy, L.-C. G. Gonzalez-Gurrola and M. Montes, *Detecting Mental Disorders in Social Media Through Emotional Patterns-the Case of Anorexia and Depression*, IEEE Transactions on Affective Computing (2021) (cit. on pp. 58, 79, 101, 102).

[324] A. Vaswani et al., *Attention Is All You Need*, Advances in Neural Information Processing Systems (2017) pp. 5998–6008 (cit. on pp. 58, 60, 76, 77, 87, 88).

[325] J. Gehring, M. Auli, D. Grangier, D. Yarats and Y. N. Dauphin, *Convolutional Sequence to Sequence Learning*, International Conference on Machine Learning (2017) pp. 1243–1252 (cit. on p. 58).

[326] J. Lee et al., *Set Transformer: A Framework for Attention-Based Permutation-Invariant Neural Networks*, International Conference on Machine Learning (2019) pp. 3744–3753 (cit. on pp. 58, 88).

[327] Y. Liu et al., *Roberta: A Robustly Optimized Bert Pretraining Approach*, arXiv preprint arXiv:1907.11692 (2019) (cit. on pp. 59, 79, 107).

[328] W. Wang et al., *Minilm: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers*, Advances in Neural Information Processing Systems 33 (2020) pp. 5776–5788 (cit. on pp. 59, 79).

[329] L. N. Smith, *Cyclical Learning Rates for Training Neural Networks*, Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV) (2017) pp. 464–472 (cit. on p. 60).

[330] D. E. Losada, F. A. Crestani and J. Parapar, *Overview of eRisk at CLEF 2019: Early Risk Prediction on the Internet (Extended Overview)*, Working Notes of CLEF 2019, Early Risk Prediction on the Internet Lab (eRisk), CEUR 2380 (2019) (cit. on p. 62).

[331] Y. Wang et al., *Self-Instruct: Aligning Language Model With Self Generated Instructions*, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2022) pp. 13484–13508 (cit. on p. 66).

[332] K. Song, X. Tan, T. Qin, J. Lu and T.-Y. Liu, *Mpnet: Masked and Permuted Pre-Training for Language Understanding*, Proceedings of the Annual Conference on Neural Information Processing Systems 33 (2020) pp. 16857–16867 (cit. on p. 66).

[333] A. Aich et al., *Towards Intelligent Clinically-Informed Language Analyses of People With Bipolar Disorder and Schizophrenia*, Findings of the Conference on Empirical Methods in Natural Language Processing (2022) pp. 2871–2887 (cit. on pp. 67, 70).

[334] D. Owen et al., *Enabling Early Health Care Intervention by Detecting Depression in Users of Web-Based Forums Using Language Models: Longitudinal Analysis and Evaluation*, JMIR AI 2 (2023) (cit. on pp. 67, 70).

[335] J. D. M.-W. C. Kenton and L. K. Toutanova, *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*, "Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (2019) pp. 4171–4186 (cit. on pp. 74, 79, 80, 107).

[336] Y. Wang, Z. Wang, C. Li, Y. Zhang and H. Wang, *A Multimodal Feature Fusion-Based Method for Individual Depression Detection on Sina Weibo*, 2020 IEEE 39th International Performance Computing and Communications Conference (IPCCC) (2020) pp. 1–8 (cit. on p. 74).

[337] I. M. Baytas et al., *Patient Subtyping via Time-Aware LSTM Networks*, Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2017) pp. 65–74 (cit. on pp. 74, 75, 79, 81).

[338] S. M. Kazemi et al., *Time2vec: Learning a Vector Representation of Time*, arXiv preprint arXiv:1907.05321 (2019) (cit. on pp. 74, 76, 79, 80).

[339] P. Dufter, M. Schmitt and H. Schütze, *Position Information in Transformers: An Overview*, Computational Linguistics (2021) pp. 1–31 (cit. on pp. 74, 77).

[340] A. Radford et al., *Learning Transferable Visual Models From Natural Language Supervision*, Proceedings of the International Conference on Machine Learning (2021) pp. 8748–8763 (cit. on pp. 76, 78).

[341] T. Kim and P. Vossen, *Emoberta: Speaker-Aware Emotion Recognition in Conversation With Roberta*, arXiv preprint arXiv:2108.12009 (2021) (cit. on pp. 76, 79).

[342] H. Tan and M. Bansal, *LXMERT: Learning Cross-Modality Encoder Representations From Transformers*, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (2019) pp. 5100–5111 (cit. on p. 76).

[343] N. Berry et al., *# WhyWeTweetMH: Understanding Why People Use Twitter to Discuss Mental Health Problems*, Journal of Medical Internet Research 19 (2017) e6173 (cit. on p. 77).

[344] J. Deng et al., *Imagenet: A Large-Scale Hierarchical Image Database*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2009) pp. 248–255 (cit. on p. 78).

[345] M. Caron et al., *Emerging Properties in Self-Supervised Vision Transformers*, Proceedings of the IEEE/CVF International Conference on Computer Vision (2021) pp. 9650–9660 (cit. on p. 78).

[346] L. Ericsson, H. Gouk and T. M. Hospedales, *How Well Do Self-Supervised Models Transfer?*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021) pp. 5414–5423 (cit. on p. 78).

[347] J. S. Lara, M. E. Aragón, F. A. González and M. Montes-y-Gómez, *Deep Bag-of-Sub-Emotions for Depression Detection in Social Media*, Proceedings of the International Conference on Text, Speech, and Dialogue (2021) pp. 60–72 (cit. on pp. 79, 101, 102).

[348] Y.-A. Wang and Y.-N. Chen, *What Do Position Embeddings Learn? An Empirical Study of Pre-Trained Language Model Positional Encoding*, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (2020) pp. 6840–6849 (cit. on pp. 79, 88).

[349] T. Gui et al.,
*Depression Detection on Social Media With Reinforcement Learning*,
China National Conference on Chinese Computational Linguistics (2019) pp. 613–624 (cit. on pp. 80, 81).

[350] C. Szegedy, S. Ioffe, V. Vanhoucke and A. A. Alemi, *Inception-V4, Inception-ResNet and the Impact of Residual Connections on Learning*,
Proceedings of the AAAI Conference on Artificial Intelligence (2017) pp. 4278–4284 (cit. on p. 80).

[351] J. Chung, C. Gulcehre, K. Cho and Y. Bengio, *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*,
Proceedings of the 2014 Workshop on Deep Learning, Annual Conference on Neural Information Processing Systems (2014) (cit. on p. 80).

[352] K. Simonyan and A. Zisserman,
*Very Deep Convolutional Networks for Large-Scale Image Recognition*,
Proceedings of the International Conference on Learning Representations (2015) (cit. on p. 80).

[353] L. N. Smith, *No More Pesky Learning Rate Guessing Games*,
arXiv preprint arXiv:1206.1106 (2015), arXiv: 1506.01186 (cit. on p. 80).

[354] S. Zanwar, D. Wiechmann, Y. Qiao and E. Kerz,
*SMHD-GER: A Large-Scale Benchmark Dataset for Automatic Mental Health Detection From Social Media in German*,
Findings of the Association for Computational Linguistics: EACL 2023 (2023) pp. 1526–1541 (cit. on p. 86).

[355] M. Ravanelli et al.,
*Multi-Task Self-Supervised Learning for Robust Speech Recognition*,
IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020) pp. 6989–6993 (cit. on pp. 86, 89).

[356] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos and M. Pantic,
*Estimation of Continuous Valence and Arousal Levels From Faces in Naturalistic Conditions*, Nature Machine Intelligence 3 (2021) pp. 42–50 (cit. on pp. 86, 89).

[357] A. Bulat and G. Tzimiropoulos,
*How Far Are We From Solving the 2d & 3d Face Alignment Problem? (And a Dataset of 230,000 3d Facial Landmarks)*,
Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017) pp. 1021–1030 (cit. on pp. 86, 89).

[358] H. Xu et al., *GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models*, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020) pp. 6183–6192 (cit. on pp. 86, 89).

[359] X. Zhang et al., *ETH-XGaze: A Large Scale Dataset for Gaze Estimation Under Extreme Head Pose and Gaze Variation*, European Conference on Computer Vision (ECCV) (2020) pp. 365–381 (cit. on pp. 86, 89).

[360] W. Zeng et al., *Real-Time Multi-Person Eyeblink Detection in the Wild for Untrimmed Video*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) pp. 13854–13863 (cit. on pp. 86, 90).

[361] M. Gales and S. Young, *The Application of Hidden Markov Models in Speech Recognition*, Now Publishers Inc, 2008 (cit. on p. 87).

[362] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*, Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (2019) (cit. on p. 87).

[363] P. Harzig, M. Einfalt and R. Lienhart, *Synchronized Audio-Visual Frames With Fractional Positional Encoding for Transformers in Video-to-Text Translation*, 2022 IEEE International Conference on Image Processing (ICIP) (2022) pp. 2041–2045 (cit. on pp. 87, 88).

[364] W. C. Sleeman IV, R. Kapoor and P. Ghosh, *Multimodal Classification: Current Landscape, Taxonomy and Future Directions*, ACM Computing Surveys 55 (2022) pp. 1–31 (cit. on p. 88).

[365] H. Bredin et al., *Pyannote.Audio: Neural Building Blocks for Speaker Diarization*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020) pp. 7124–7128 (cit. on p. 89).

[366] H. Bredin and A. Laurent, *End-to-End Speaker Segmentation for Overlap-Aware Resegmentation*, Proceedings of Interspeech (2021) pp. 3111–3115 (cit. on p. 89).

[367] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte and Y. Bengio, *Learning Problem-Agnostic Speech Representations From Multiple Self-Supervised Tasks*, Proceedings of Interspeech (2019) pp. 161–165 (cit. on p. 89).

[368] Z. Fang et al., *Learning Coordinated Emotion Representation Between Voice and Face*, Applied Intelligence 53 (2023) pp. 14470–14492 (cit. on p. 89).

[369] J. A. Russell, *A Circumplex Model of Affect.*, Journal of Personality and Social Psychology 39 (1980) pp. 1161 (cit. on p. 89).

[370] H. Gunes and B. Schuller, *Categorical and Dimensional Affect Analysis in Continuous Input: Current Trends and Future Directions*, Image and Vision Computing 31 (2013) pp. 120–136 (cit. on p. 89).

[371] V. Bazarevsky et al., *Blazepose: On-Device Real-Time Body Pose Tracking*, Proceedings of the CV4ARVR Workshop, CVPR (2020) (cit. on p. 89).

[372] H. Dibeklioğlu, Z. Hammal and J. F. Cohn, *Dynamic Multimodal Measurement of Depression Severity Using Deep Autoencoding*, IEEE Journal of Biomedical and Health Informatics 22 (2017) pp. 525–536 (cit. on p. 90).

[373] M. Valstar et al., *Avec 2013: The Continuous Audio/Visual Emotion and Depression Recognition Challenge*, Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge (2013) pp. 3–10 (cit. on p. 90).

[374] S. Alghowinem et al., *From Joyous to Clinically Depressed: Mood Detection Using Spontaneous Speech*, Proceedings of the International Florida Artificial Intelligence Research Society Conference 19 (2012) pp. 141–146 (cit. on p. 90).

[375] G. Degottex, J. Kane, T. Drugman, T. Raitio and S. Scherer, *COVAREP — A Collaborative Voice Analysis Repository for Speech Technologies*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2014) pp. 960–964 (cit. on p. 90).

[376] T. Baltrušaitis, P. Robinson and L.-P. Morency, *OpenFace: An Open Source Facial Behavior Analysis Toolkit*, Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV) (2016) pp. 1–10 (cit. on pp. 90, 91).

[377] S. Davis and P. Mermelstein, *Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences*, IEEE Transactions on Acoustics, Speech, and Signal Processing 28 (1980) pp. 357–366 (cit. on p. 90).

[378] F. Eyben et al., *The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing*, IEEE Transactions on Affective Computing 7 (2016) pp. 190–202 (cit. on p. 90).

[379] K. He, X. Zhang, S. Ren and J. Sun, *Deep Residual Learning for Image Recognition*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2016) pp. 770–778 (cit. on pp. 90, 91).

[380] G. Coppersmith, R. Leary, P. Crutchley and A. Fine, *Natural Language Processing of Social Media as Screening for Suicide Risk*, Biomedical Informatics Insights 10 (2018) pp. 1178222618792860 (cit. on p. 92).

[381] R. Kaushik, S. Gaur, J. N. Pandit, S. Satapathy and C. Behera, *Live Streaming of Suicide on Facebook*, Psychiatry Research Case Reports 2 (2023) pp. 100141 (cit. on p. 92).

[382] I.Loshchilov and F. Hutter, *Decoupled Weight Decay Regularization*, Proceedings of the International Conference on Learning Representations (2019) (cit. on p. 92).

[383] C. Doersch, A. Gupta and A. Zisserman, *Crosstransformers: Spatially-Aware Few-Shot Transfer*, Advances in Neural Information Processing Systems 33 (2020) pp. 21981–21993 (cit. on p. 93).

[384] B. Li, P. Xiong, C. Han and T. Guo, *Shrinking Temporal Attention in Transformers for Video Action Recognition*, Proceedings of the AAAI Conference on Artificial Intelligence 36 (2022) pp. 1263–1271 (cit. on p. 93).

[385] J. R. Williamson et al., *Detecting Depression Using Vocal, Facial and Semantic Communication Cues*, Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, AVEC '16 (2016) pp. 11–18 (cit. on pp. 93, 94, 97, 127).

[386] S. Song, L. Shen and M. Valstar, *Human Behaviour-Based Automatic Depression Analysis Using Hand-Crafted Statistics and Deep Learned Spectral Features*, 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG) (2018) pp. 158–165 (cit. on pp. 93, 94, 97, 127).

[387] P.-C. Wei et al., *Multi-Modal Depression Estimation Based on Sub-Attentional Fusion*, Proceedings of the European Conference on Computer Vision (2023) pp. 623–639 (cit. on pp. 93, 94, 97, 127).

[388] X. Ma, H. Yang, Q. Chen, D. Huang and Y. Wang, *DepAudioNet: An Efficient Deep Model for Audio Based Depression Classification*, Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, AVEC '16 (2016) pp. 35–42 (cit. on pp. 93, 94, 97, 127).

[389] E. Villatoro-Tello, G. Ramírez-de-la-Rosa, D. Gática-Pérez, M. Magimai.-Doss and H. Jiménez-Salazar, *Approximating the Mental Lexicon From Clinical Interviews as a Support Tool for Depression Detection*, Proceedings of the 2021 International Conference on Multimodal Interaction (2021) pp. 557–566 (cit. on p. 93).

[390] G. S. Saggu, K. Gupta, K. Arya and C. R. Rodriguez, *DepressNet: A Multimodal Hierarchical Attention Mechanism Approach for Depression Detection*, International Journal of Engineering Sciences 15 (2022) pp. 24–32 (cit. on p. 93).

[391] A. Ray, S. Kumar, R. Reddy, P. Mukherjee and R. Garg, *Multi-Level Attention Network Using Text, Audio and Video for Depression Prediction*, Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (2019) pp. 81–88 (cit. on p. 93).

[392] A. Jaegle et al., *Perceiver: General Perception With Iterative Attention*, International Conference on Machine Learning (2021) pp. 4651–4664 (cit. on pp. 93–95).

[393] Y. Bengio, P. Simard and P. Frasconi, *Learning Long-Term Dependencies With Gradient Descent Is Difficult*, IEEE Transactions on Neural Networks 5 (1994) pp. 157–166 (cit. on p. 93).

[394] R. Pascanu, T. Mikolov and Y. Bengio, *On the Difficulty of Training Recurrent Neural Networks*, International Conference on Machine Learning (2013) pp. 1310–1318 (cit. on p. 93).

[395] K. Jaidka et al., *Beyond Positive Emotion: Deconstructing Happy Moments Based on Writing Prompts*, Proceedings of the International AAAI Conference on Web and Social Media (2020) (cit. on pp. 99, 101, 102, 107, 110, 119, 120, 122).

[396] A. Waterman, *Two Conceptions of Happiness: Contrasts of Personal Expressiveness (Eudaimonia) and Hedonic Enjoyment*, Journal of Personality and Social Psychology 64 (1993) (cit. on p. 99).

[397] A. Moreno-Ortiz, C. Pérez-Hernández and M. García-Gámez, *The Language of Happiness in Self-Reported Descriptions of Happy Moments: Words, Concepts, and Entities*, Humanities and Social Sciences Communications 9 (2022) pp. 1–13 (cit. on pp. 99, 101, 105, 119, 122).

[398] A. Asai et al., *HappyDB: A Corpus of 100,000 Crowdsourced Happy Moments*, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (2018) (cit. on pp. 99, 101, 103–105, 107, 110, 119, 122).

[399] C. Elkan and K. Noto, *Learning Classifiers From Only Positive and Unlabeled Data*, Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2008) pp. 213–220 (cit. on pp. 99, 106, 107).

[400] D. L. Paulhus and P. D. Trapnell, *Self-Presentation of Personality: An Agency-Communion Framework.*, The Guilford Press (2008) (cit. on pp. 99, 101, 110).

[401] K. Jaidka, S. Mumick, N. Chhaya and L. Ungar,
*The CL-Aff Happiness Shared Task: Results and Key Insights*,
Proceedings of the Workshop on Affective Content Analysis (2019)
(cit. on pp. 99–101, 105, 110, 120).

[402] P. Spinhoven, B. M. Elzinga, E. Giltay and B. W. Penninx,
*Anxious or Depressed and Still Happy?*, PloS One 10 (2015) e0139912
(cit. on pp. 100, 101).

[403] A. Tigunova, A. Yates, P. Mirza and G. Weikum, *Listening Between the Lines: Learning Personal Attributes From Conversations*, Proceedings of the ACM World Wide Web Conference (2019) pp. 1818–1828
(cit. on pp. 101, 112, 113, 122).

[404] R. Mihalcea and H. Liu, *A Corpus-Based Approach to Finding Happiness.*,
AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs (2006) (cit. on p. 101).

[405] C. Lin et al., *Sensemood: Depression Detection on Social Media*,
Proceedings of the ACM International Conference on Multimedia Retrieval (2020) pp. 407–411 (cit. on p. 101).

[406] D. Ramírez-Cifuentes et al., *Detection of Suicidal Ideation on Social Media: Multimodal, Relational, and Behavioral Analysis*,
Journal of Medical Internet Research 22 (2020) e17758 (cit. on p. 101).

[407] A. G. Reece et al.,
*Forecasting the Onset and Course of Mental Illness With Twitter Data*,
Scientific Reports 7 (2017) pp. 13006 (cit. on p. 101).

[408] S. G. Burdisso, M. Errecalde and M. Montes-y-Gómez,
*A Text Classification Framework for Simple and Effective Early Depression Detection Over Social Media Streams*,
Expert systems with applications 133 (2019) pp. 182–197 (cit. on p. 101).

[409] D. Bakan, *The Duality of Human Existence: Isolation and Communion in Western Man*, Beacon Press, 1966 (cit. on p. 102).

[410] E. L. Deci and R. M. Ryan,
*Intrinsic Motivation and Self-Determination in Human Behavior*,
Springer Science & Business Media, 2013 (cit. on p. 102).

[411] R. M. Ryan and E. L. Deci, *Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being.*,
American Psychologist 55 (2000) pp. 68 (cit. on p. 102).

[412] R. M. Ryan,
*Psychological Needs and the Facilitation of Integrative Processes*,
Journal of Personality 63 (1995) pp. 397–427 (cit. on p. 102).

[413] R. M. Ryan and E. L. Deci, *On Happiness and Human Potentials: A Review of Research on Hedonic and Eudaimonic Well-Being*, Annual Review of Psychology 52 (2001) pp. 141–166 (cit. on p. 102).

[414] R. DeCharms, *Personal Causation: The International Affective Determinations of Behavior*, Acad. Press, 1968 (cit. on p. 102).

[415] R. W. White, *Motivation Reconsidered: The Concept of Competence.*, Psychological Review 66 (1959) pp. 297 (cit. on p. 102).

[416] V. S. Helgeson, *Relation of Agency and Communion to Well-Being: Evidence and Potential Explanations.*, Psychological Bulletin 116 (1994) pp. 412 (cit. on p. 102).

[417] C. Saragovi, J. Aubé, R. Koestner and D. Zuroff, *Traits, Motives, and Depressive Styles as Reflections of Agency and Communion*, Personality and Social Psychology Bulletin 28 (2002) pp. 563–577 (cit. on pp. 102, 121).

[418] C. P. Keeton, M. Perry-Jenkins and A. G. Sayer, *Sense of Control Predicts Depressive and Anxious Symptoms Across the Transition to Parenthood.*, Journal of Family Psychology 22 (2008) pp. 212 (cit. on p. 102).

[419] P. Gilbert, *Depression: The Evolution of Powerlessness*, Routledge, 2016 (cit. on p. 102).

[420] S. L. Bem, *The Measurement of Psychological Androgyny.*, Journal of Consulting and Clinical Psychology 42 (1974) pp. 155 (cit. on p. 103).

[421] J. T. Spence, R. Helmreich and J. Stapp, *Ratings of Self and Peers on Sex Role Attributes and Their Relation to Self-Esteem and Conceptions of Masculinity and Femininity.*, Journal of Personality and Social Psychology 32 (1975) pp. 29 (cit. on p. 103).

[422] J. T. Spence and R. L. Helmreich, *Masculinity and Femininity: Their Psychological Dimensions, Correlates, and Antecedents*, University of Texas Press, 1978 (cit. on p. 103).

[423] A. H. Eagly, *Sex Differences in Social Behavior: A Social-Role Interpretation*, Lawrence Erlbaum, 1987 (cit. on p. 103).

[424] W. Wood and A. H. Eagly, "Biosocial Construction of Sex Differences and Similarities in Behavior", vol. 46, Elsevier, 2012 pp. 55–123 (cit. on p. 103).

[425] N. Hsu, K. L. Badura, D. A. Newman and M. E. P. Speach, *Gender,"masculinity," and "Femininity": A Meta-Analytic Review of Gender Differences in Agency and Communion.*, Psychological Bulletin 147 (2021) pp. 987 (cit. on pp. 103, 121, 129).

[426] D. McDuff and S. Glass, *Faded Smiles? A Largescale Observational Study of Smiling From Adolescence to Old Age*, IEEE Transactions on Affective Computing 12 (2019) pp. 53–65 (cit. on p. 103).

[427] D. McDuff, E. Jun, K. Rowan and M. Czerwinski, *Longitudinal Observational Evidence of the Impact of Emotion Regulation Strategies on Affective Expression*, IEEE Transactions on Affective Computing 12 (2019) pp. 636–647 (cit. on p. 103).

[428] R. Al-Rfou', B. Perozzi and S. Skiena, *Polyglot: Distributed Word Representations for Multilingual NLP*, Proceedings of the Seventeenth Conference on Computational Natural Language Learning (2013) pp. 183–192 (cit. on p. 105).

[429] J. Bekker and J. Davis, *Learning From Positive and Unlabeled Data: A Survey*, Machine Learning 109 (2020) pp. 719–760 (cit. on p. 106).

[430] F. Mordelet and J.-P. Vert, *Prodige: Prioritization of Disease Genes With Multitask Machine Learning From Positive and Unlabeled Examples*, BMC Bioinformatics 12 (2011) pp. 1–15 (cit. on p. 106).

[431] F. Mordelet and J.-P. Vert, *A Bagging SVM to Learn From Positive and Unlabeled Examples*, Pattern Recognition Letters 37 (2014) pp. 201–209 (cit. on p. 106).

[432] J. Yi, C.-J. Hsieh, K. R. Varshney, L. Zhang and Y. Li, *Scalable Demand-Aware Recommendation*, Proceedings of the Annual Conference on Neural Information Processing Systems 30 (2017) (cit. on p. 106).

[433] D. Hernández-Fusilier, M. Montes-y-Gómez, P. Rosso and R. Guzmán-Cabrera, *Detecting Positive and Negative Deceptive Opinions Using PU-learning*, Information Processing & Management 51 (2015) pp. 433–443 (cit. on p. 106).

[434] Y. Ren, D. Ji and H. Zhang, *Positive Unlabeled Learning for Deceptive Reviews Detection*, Proceedings of the Conference on Empirical Methods in Natural Language Processing (2014) pp. 488–498 (cit. on p. 106).

[435] F. He, T. Liu, G. I. Webb and D. Tao, *Instance-Dependent Pu Learning by Bayesian Optimal Relabeling*, arXiv preprint arXiv:1808.02180 (2018) (cit. on p. 106).

[436] R. Kamath, A. Ghoshal, S. Eswaran and P. Honnavalli, *An Enhanced Context-Based Emotion Detection Model Using RoBERTa*, Proceedings of the IEEE international conference on electronics, computing and communication technologies (2022) pp. 1–6 (cit. on p. 107).

[437] D. Demszky et al., *GoEmotions: A Dataset of Fine-Grained Emotions*, Proceedings of the Annual Meeting of the Association for Computational Linguistics (2020) pp. 4040–4054 (cit. on p. 107).

[438] Q. Xie, Z. Dai, E. Hovy, T. Luong and Q. Le, *Unsupervised Data Augmentation for Consistency Training*, Proceedings of the Annual Conference on Neural Information Processing Systems 33 (2020) pp. 6256–6268 (cit. on p. 111).

[439] J. Tiedemann and S. Thottingal, *Opus-Mt — Building open translation services for the world*, Proceedings of the Annual Conference of the European Association for Machine Translation (2020) (cit. on p. 111).

[440] A. Rajendran, C. Zhang and M. Abdul-Mageed, *Happy Together: Learning and Understanding Appraisal From Natural Language*, Proceedings of the Workshop on Affective Content Analysis (2019) (cit. on pp. 111, 112).

[441] B. Syed, V. Indurthi, K. Shah, M. Gupta and V. Varma, *Ingredients for Happiness: Modeling Constructs via Semi-Supervised Content Driven Inductive Transfer Learning*, Proceedings of the Workshop on Affective Content Analysis (2019) (cit. on pp. 111, 112).

[442] M. Saxon, S. Bhandari, L. Ruskin and G. Honda, *Word Pair Convolutional Model for Happy Moment Classification.*, Proceedings of the Workshop on Affective Content Analysis (2019) (cit. on pp. 111, 112).

[443] B. Talafha and M. Al-Ayyoub, *Ioh-Rcnn: Pursuing the Ingredients of Happiness Using Recurrent Convolutional Neural Networks*, Proceedings of the Workshop on Affective Content Analysis (2019) (cit. on pp. 111, 112).

[444] J. Wu et al., *Cruzaffect at Affcon 2019 Shared Task: A Feature-Rich Approach to Characterize Happiness*, Proceedings of the Workshop on Affective Content Analysis (2019) (cit. on pp. 111, 112).

[445] M. Grootendorst, *BERTopic: Neural Topic Modeling With a Class-Based TF-IDF Procedure*, arXiv preprint arXiv:2203.05794 (2022) (cit. on p. 112).

[446] D. M. Blei, A. Y. Ng and M. I. Jordan, *Latent Dirichlet Allocation*, Journal of Machine Learning Research 3 (2003) pp. 993–1022 (cit. on p. 113).

[447] S. Deerwester, S. Dumais, G. Furnas, T. Landauer and R. Harshman, *Indexing by Latent Semantic Analysis.*, Journal of the American Society for Information Science 41, 391–407 (1990) (cit. on p. 113).

[448] L. McInnes, J. Healy, N. Saul and L. Großberger, *UMAP: Uniform Manifold Approximation and Projection*, Journal of Open Source Software 3 (2018) pp. 861 (cit. on p. 113).

[449] A. Tigunova, P. Mirza, A. Yates and G. Weikum, *RedDust: A Large Reusable Dataset of Reddit User Traits*, Proceedings of the Twelfth Language Resources and Evaluation Conference (2020) pp. 6118–6126 (cit. on pp. 113, 114, 122).

[450] C. Aguirre, K. Harrigian and M. Dredze, *Gender and Racial Fairness in Depression Research Using Social Media*, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (2021) pp. 2932–2949 (cit. on pp. 113, 122).

[451] I. Lin et al., *Gendered Mental Health Stigma in Masked Language Models*, Proceedings of the Conference on Empirical Methods in Natural Language Processing (2022) pp. 2152–2170 (cit. on pp. 113, 122).

[452] G. Jagfeld, F. Lobban, P. Rayson and S. Jones, *Understanding Who Uses Reddit: Profiling individuals with a self-reported bipolar disorder diagnosis*, Proceedings of the Workshop on Computational Linguistics and Clinical Psychology (CLPsych) (2021) pp. 1–14 (cit. on p. 113).

[453] A. Amaya, R. Bach, F. Keusch and F. Kreuter, *New Data Sources in Social Science Research: Things to Know Before Working With Reddit Data*, Social Science Computer Review 39 (2021) pp. 943–960 (cit. on p. 115).

[454] C. Kuehner, *Why Is Depression More Common Among Women Than Among Men?*, Lancet Psychiatry 4 (2017) pp. 146–158 (cit. on p. 115).

[455] T. Elmer and C. Stadtfeld, *Depressive Symptoms Are Associated With Social Isolation in Face-to-Face Interaction Networks*, Scientific Reports 10 (2020) pp. 1444 (cit. on p. 120).

[456] L. H. Brown, T. Strauman, N. Barrantes-Vidal, P. J. Silvia and T. R. Kwapil, *An Experience-Sampling Study of Depressive Symptoms and Their Social Context*, Journal of Nervous and Mental Disease 199 (2011) pp. 403–409 (cit. on p. 120).

[457] E. S. Chevron, D. M. Quinlan and S. J. Blatt,
*Sex Roles and Gender Differences in the Experience of Depression.*,
Journal of Abnormal Psychology 87 (1978) pp. 680 (cit. on p. 120).

[458] D. D. Logan, H. N. College and E. Kaschak,
*The Relationship of Sex, Sex Role, and Mental Health*,
Psychology of Women Quarterly 4 (1980) pp. 573–580 (cit. on p. 120).

[459] R. J. Waldinger and M. S. Schulz,
*What's Love Got to Do With It? Social Functioning, Perceived Health, and Daily Happiness in Married Octogenarians.*,
Psychology and Aging 25 (2010) pp. 422 (cit. on pp. 121, 122).

[460] J. F. Helliwell and R. D. Putnam, *The Social Context of Well–Being*,
Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 359 (2004) pp. 1435–1446 (cit. on p. 121).

[461] M. Haller and M. Hadler,
*How Social Relations and Structures Can Produce Happiness and Unhappiness: An International Comparative Analysis*,
Social Indicators Research 75 (2006) pp. 169–216 (cit. on p. 121).

[462] V. Amati, S. Meggiolaro, G. Rivellini and S. Zaccarin,
*Social Relations and Life Satisfaction: The Role of Friends*,
Genus 74 (2018) pp. 1–18 (cit. on p. 121).

[463] A. R. Teo, H. Choi and M. Valenstein, *Social Relationships and Depression: Ten-Year Follow-Up From a Nationally Representative Study*,
PloS One 8 (2013) e62396 (cit. on p. 121).

[464] K. McTiernan, F. Gullon-Scott and R. Dudley,
*Do Positive Psychology Interventions Impact on the Subjective Wellbeing and Depression of Clients? A Systematic Methodological Review*,
Journal of Contemporary Psychotherapy 52 (2022) pp. 1–13
(cit. on p. 121).

[465] S. Evensen et al.,
*Happiness Entailment: Automating Suggestions for Well-Being*,
Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction (2019) pp. 62–68 (cit. on p. 123).

[466] A. Benton, G. Coppersmith and M. Dredze,
*Ethical Research Protocols for Social Media Health Research*,
Proceedings of the First ACL Workshop on Ethics in Natural Language Processing (2017) pp. 94–102 (cit. on pp. 127, 128).

[467] S. Chancellor, M. L. Birnbaum, E. D. Caine, V. M. Silenzio and M. De Choudhury, *A Taxonomy of Ethical Tensions in Inferring Mental Health States From Social Media*, Proceedings of the Conference on Fairness, Accountability, and Transparency (2019) pp. 79–88
(cit. on pp. 127–129).

[468] D. Hovy and S. L. Spruit,
*The Social Impact of Natural Language Processing*,
Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (2016) pp. 591–598 (cit. on p. 128).

[469] L. J. Kirmayer, *Cultural variations in the clinical presentation of depression and anxiety: implications for diagnosis and treatment*,
Journal of Clinical Psychiatry 62 (2001) pp. 22–30 (cit. on p. 128).

[470] R. S. Tokunaga, *High-speed internet access to the other: The influence of cultural orientations on self-disclosures in offline and online relationships*,
Journal of Intercultural Communication Research 38 (2009) pp. 133–147 (cit. on p. 128).

[471] K. M. Broczek, M.-C. Gely-Nargeot and P. Gareri,
*Editorial: Depression across cultures and linguistic identities*,
Frontiers in Psychology 15 (2024) (cit. on p. 128).

[472] I. H. Gotlib and C. L. Hammen, *Handbook of depression*, (2008) (cit. on p. 128).

[473] V. N. S. De Snyder, M. de Jesus Diaz-Perez and V. D. Ojeda,
*The prevalence of nervios and associated symptomatology among inhabitants of Mexican rural communities*,
Culture, Medicine and Psychiatry 24 (2000) pp. 453–470 (cit. on p. 128).

[474] S. R. Pendse, K. Niederhoffer and A. Sharma,
*Cross-cultural differences in the use of online mental health support forums*,
Proceedings of the ACM on Human-Computer Interaction 3 (2019) pp. 1–29 (cit. on p. 128).

[475] J. Charteris-Black,
*Shattering the bell jar: Metaphor, gender, and depression*,
Metaphor and Symbol 27 (2012) pp. 199–216 (cit. on p. 129).

[476] S. Chancellor, E. P. Baumer and M. De Choudhury,
*Who Is The" Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health From Social Media*, Proceedings of the ACM on Human-Computer Interaction 3 (2019) pp. 1–32 (cit. on pp. 129, 130).

[477] M. Piccinelli and G. Wilkinson,
*Gender Differences in Depression: Critical Review*,
British Journal of Psychiatry 177 (2000) pp. 486–492 (cit. on p. 129).

[478] R. H. Salk, J. S. Hyde and L. Y. Abramson,
*Gender Differences in Depression in Representative National Samples: Meta-Analyses of Diagnoses and Symptoms.*,
Psychological Bulletin 143 (2017) pp. 783 (cit. on p. 129).

[479] E. Bao, A. Pérez and J. Parapar,
*Explainable Depression Symptom Detection in Social Media*,
Health Information Science and Systems 12 (2024) pp. 47 (cit. on p. 130).