

Explainable Deep Learning for the Early Detection of Mental Health Risks in Social Media

Hessam Amini

**A Thesis
in
The Department
of
Computer Science and Software Engineering**

**Presented in Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy (Computer Science) at
Concordia University
Montréal, Québec, Canada**

December 2025

© Hessam Amini, 2026

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Hessam Amini**

Entitled: **Explainable Deep Learning for the Early Detection of Mental Health
Risks in Social Media**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to
originality and quality.

Signed by the Final Examining Committee:

Dr. Frédéric Godin Chair

Dr. Diana Inkpen External Examiner

Dr. Emad Shihab Examiner

Dr. Marta Kersten-Oertel Examiner

Dr. Ferhat Khendek Examiner

Dr. Leila Kosseim Supervisor

Approved by

Dr. Joey Paquet, Chair
Department of Computer Science and Software Engineering

2025

Dr. Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Explainable Deep Learning for the Early Detection of Mental Health Risks in Social Media

Hessam Amini, Ph.D.

Concordia University, 2026

This thesis explores explainable deep learning methods for the early detection of mental health risks from social media posts. We focus on models that not only detect risk early, but also provide clear post-level explanations that align with their decision process.

First, we examine baseline architectures for user-level early detection of anorexia within the CLEF eRisk framework and report on our system’s performance in the 2019 shared task. Our submission combined attention-based neural sub-models with a final Support Vector Machine (SVM) classifier and ranked first on the official metrics, with an F1 of 0.7073 and a latency-weighted F1 of 0.6908. An analysis of the user-level attention as post-level evidence shows that a small subset of posts often drives the final decision.

Second, we introduce ATTEFA, a quantitative metric for attention faithfulness based on a single-run adversarial setup. The method trains an adversarial model to match the base model’s predictions while diverging in attention, allowing us to measure how much attention can change without affecting outputs. Across synthetic and real datasets, ATTEFA reveals when attention aligns with the model’s decision process and when it does not.

Finally, we present a user-aware attention framework built on DISORBERT and fine-tuned end-to-end. We pair this model with adversarial training and evaluate faithfulness using ATTEFA. On four eRisk tasks (depression, anorexia, self-harm, and pathological gambling), the approach maintains competitive latency-weighted F1 scores while producing transparent post-level explanations. Our analysis further shows that decisions typically rely on a small portion of each user’s timeline, highlighting both the sparsity of strong mental health signals and the value of faithful explanations.

Taken together, these contributions build a step-by-step path: from strong early detection, to a principled measure of attention faithfulness, to a practical, user-aware model whose explanations are both transparent and aligned with its decision process.

Acknowledgments

I would first like to express my deepest gratitude to my supervisor, Dr. Leila Kosseim, for her continuous support, patience, and encouragement throughout my Ph.D. studies. Her guidance and understanding, especially during the long duration of this journey, have been invaluable. I could not have asked for a better mentor and role model.

I would also like to thank my thesis committee members, Dr. Emad Shihab, Dr. Marta Kersten-Oertel, Dr. Ferhat Khendek, and Dr. Weiyi (Ian) Shang, as well as my external examiner, Dr. Diana Inkpen, for their constructive feedback and insightful comments, which greatly improved the quality of this work.

I am forever grateful to my family (my father, mother, and brother) for their constant mentorship and support. They stood by me unconditionally, supporting me emotionally and financially through my studies. Their sacrifices and encouragement made this achievement possible.

My deepest thanks go to my spouse, Elham. Since the beginning of our life together, she has been my constant source of strength and inspiration. Her unwavering support, motivation, and belief in me carried me through every challenge of this journey. Beyond her emotional support, her scientific perspective and thoughtful guidance played an important role in shaping this work. I cannot imagine reaching this milestone without her by my side.

Finally, I would like to dedicate this thesis to all those who continue to inspire me with their resilience and strength, and who remind me of the importance of striving for knowledge, empathy, and justice.

Contribution of Authors

The main chapters in this dissertation (Chapters 3, 4, and 5) are based on papers already published. Below is a description of our individual contributions.

- Chapter 3 on early detection of anorexia in social media has been published as an invited extended book chapter, [Amini, Mohammadi, and Kosseim \(2022\)](#):

Amini, H., Mohammadi, E., & Kosseim, L. (2022, September). Quick and (Maybe not so) Easy Detection of Anorexia in Social Media: To Explainability and Beyond. In *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the eRisk Project* (pp. 141–158). Springer International Publishing.

This chapter consolidates and extends our earlier publications: the CLEF eRisk 2019 shared-task participation paper ([Mohammadi, Amini, & Kosseim, 2019b](#)), and our follow-up paper in the proceedings of NLDB 2020 ([Amini & Kosseim, 2020](#)).

Contributions of Authors: For the eRisk 2019 shared task paper ([Mohammadi et al., 2019b](#)), I designed the overall framework and ensemble architecture (attention-based neural sub-models and fusion+SVM). Elham Mohammadi implemented the shared-task approach, ran the experiments, and prepared the submissions. Both authors contributed equally to the writing of the paper. The subsequent NLDB 2020 paper ([Amini & Kosseim, 2020](#)), which focused on post-hoc explainability evaluation and analysis, was conducted and written solely by myself. For the invited book chapter ([Amini et al., 2022](#)), I compiled and extended the shared-task and NLDB papers and wrote the majority of the chapter. Elham Mohammadi reviewed the draft and provided feedback. Dr. Leila Kosseim contributed to the editing of all three publications and provided overall guidance and technical supervision.

- Chapter 4 on quantifying attention faithfulness has been published as [Amini and Kosseim \(2022\)](#):

Amini, H., & Kosseim, L. (2022, December). How (Un)Faithful is Attention? In *Proceedings of the 2022 Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackBoxNLP 2022)* (pp. 119–130). Abu Dhabi, UAE. (Co-located with EMNLP 2022)

Contributions of Authors: I led the project, proposed the single-run adversarial setup and the ATTEFA metric, implemented all experiments (synthetic and real-world), analyzed results, and wrote the paper. Dr. Leila Kosseim contributed to framing the problem, provided feedback on the methodology and analyses, and edited the paper.

- Chapter 5 on user-aware attention for explainable early risk detection has been published as [Amini and Kosseim \(2025\)](#):

Amini, H., & Kosseim, L. (2025, September). From Posts to Predictions: A User-Aware Framework for Faithful and Transparent Detection of Mental Health Risks on Social Media. In *Proceedings of the 2025 International Conference on Recent Advances in Natural Language Processing (RANLP 2025)* (pp. 75–84). Varna, Bulgaria.

Contributions of Authors: I led the project, designed the architecture (DISORBERT-based post encoder with user-level attention), implemented the system and evaluated attention’s faithfulness, conducted experiments and ablations on the eRisk benchmarks, and wrote the paper. Dr. Leila Kosseim contributed to editing and provided general guidance and technical direction.

Contents

List of Figures	xi
List of Tables	xii
Acronyms	xiv
1 Introduction	1
1.1 Motivation	1
1.2 Problem Statement	2
1.3 Research Objectives	3
1.4 Overall Methodology	4
1.5 Contributions	5
1.6 Thesis Organization	6
2 Literature Review	7
2.1 Introduction	7
2.2 Background	8
2.3 Deep Learning for Mental Health Detection	9
2.4 Explainability in Natural Language Processing (NLP) Models	11
2.5 Conclusion	15
3 Early Detection of Anorexia in Social Media	16
3.1 Introduction	17
3.2 Related Work	18
3.3 Task and Dataset	20

3.4	Evaluation Metrics	21
3.5	System Overview	23
3.6	Experimental Setup	27
3.7	Shared Task Results	29
3.8	Explainability Analysis	30
3.9	Conclusion	37
4	Quantitative Measurement of Attention Faithfulness	39
4.1	Introduction	40
4.2	Related Work	41
4.3	Method	43
4.4	Datasets	46
4.5	Experimental Setup	49
4.6	Results and Discussion	50
4.7	Conclusion	57
5	User-Aware Attention for Explainable Detection of Mental Health Risks in Social Media	59
5.1	Introduction	60
5.2	Related Work	61
5.3	Methodology	63
5.4	Results and Discussion	70
5.5	A Closer Look at the Attention Weights	72
5.6	Limitations	75
5.7	Conclusion	76
6	Conclusion and Future Work	77
6.1	Contributions	77
6.2	Limitations	78
6.3	Future Work	81
	Bibliography	84

Appendix A Synthetic Datasets and Adversarial Results	99
A.1 Samples from the <i>Mock</i> -* Datasets	99
A.2 All Results from the Adversarial Setup	108
Appendix B Extended Early-Prediction Evaluation Results	111

List of Figures

Figure 3.1	Model architecture overview	24
Figure 3.2	Performance vs. number of posts	33
Figure 4.1	ATTEFA distributions by model/dataset	51
Figure 4.2	JSD/TVD comparison of adversarial setups	56
Figure 5.1	User-aware attention model	64
Figure 5.2	Top-post removals to flip decisions	74

List of Tables

Table 3.1	eRisk 2019 Task 1 dataset statistics	21
Table 3.2	Sub-model hyperparameters	28
Table 3.3	eRisk 2019 Task 1 results	29
Table 3.4	Performance vs. number of posts used	32
Table 3.5	Average attention weights by post rank	34
Table 4.1	Dataset statistics	49
Table 4.2	Base model F1 scores on test datasets	51
Table 4.3	Adversarial results (avg $\pm$ std)	52
Table 5.1	eRisk dataset sources	69
Table 5.2	eRisk dataset statistics	69
Table 5.3	Base results on test datasets	70
Table 5.4	Model rankings (based on l-w F1)	71
Table 5.5	Adversarial training results	71
Table 5.6	Top-post removals to flip decisions	74
Table A.1	Mock-1 dataset samples	100
Table A.2	Mock-2 dataset samples	101
Table A.3	Mock-5 dataset samples	102
Table A.4	Mock-10 dataset samples	103
Table A.5	Mock-1q dataset samples	104
Table A.6	Mock-2q dataset samples	105
Table A.7	Mock-3q dataset samples	106
Table A.8	Mock-4q dataset samples	107
Table A.9	Adversarial results on synthetic datasets	109

Table A.10 Adversarial results on real-life datasets	110
Table B.1 Decision-based metrics of the early evaluation	112
Table B.2 Ranking-based metrics of the early evaluation	112

Acronyms

BDI Beck’s Depression Inventory. [10](#), [15](#)

BERT Bidirectional Encoder Representations from Transformers. [9](#), [10](#), [61](#), [62](#), [64](#)

CBOW Continuous Bag of Words. [50](#)

CNN Convolutional Neural Network. [4](#), [9](#), [15](#), [20](#), [25](#), [28–31](#), [35–37](#), [58](#), [63](#)

ERDE Early Risk Detection Error. [19–22](#), [29](#), [111](#), [112](#)

GRU Gated Recurrent Unit. [25](#), [28](#), [58](#)

HAN Hierarchical Attention Network. [9](#), [62](#)

JSD Jensen-Shannon Divergence. [xi](#), [42](#), [44–46](#), [51](#), [52](#), [55–57](#), [67–69](#), [109](#), [110](#)

KLD Kullback-Leibler Divergence. [44](#), [45](#), [67](#)

LDA Latent Dirichlet Allocation. [18](#)

LIME Local Interpretable Model-agnostic Explanations. [12](#), [82](#)

LIWC Linguistic Inquiry and Word Count. [19](#)

LLM Large Language Model. [8](#), [10](#), [11](#), [15](#), [81](#)

LSTM Long Short-Term Memory. [9](#), [25](#), [28](#), [41](#), [43](#), [49–58](#), [109](#), [110](#)

MLP Multi-Layer Perceptron. [41](#), [43](#), [49–56](#), [58](#), [109](#), [110](#)

NDCG Normalized Discounted Cumulative Gain. [112](#)

NLP Natural Language Processing. [viii](#), [2](#), [3](#), [6–9](#), [11](#), [12](#), [15](#), [17](#), [18](#), [40](#), [57](#), [58](#), [60](#)

PR-AUC Precision-Recall Area Under the Curve. [66](#), [69](#), [70](#)

PTSD Post-Traumatic Stress Disorder. [8](#), [18](#), [19](#), [79](#)

RNN Recurrent Neural Network. [4](#), [9](#), [15](#), [25](#), [28](#), [63](#)

ROC-AUC Receiver Operating Characteristic Area Under the Curve. [43](#)

SHAP SHapley Additive exPlanations. [12](#), [13](#), [82](#)

SVM Support Vector Machine. [iii](#), [9](#), [16](#), [18](#), [19](#), [23](#), [24](#), [27](#), [28](#)

TVD Total Variation Distance. [xi](#), [42–46](#), [51–53](#), [56](#), [57](#), [109](#), [110](#)

Chapter 1

Introduction

1.1 Motivation

Mental health is a crucial component of overall well-being, yet mental disorders are highly prevalent and commonly under-treated. Globally, around one in eight people (≈ 970 million) were living with a mental disorder in 2019, and depression alone affected about 280 million people ([World Health Organization, 2022a, 2022b](#)). In the United States, more than one in five adults experiences mental illness in a given year ([National Institute of Mental Health, 2025](#)). Evidence and global guidance emphasize the importance of early identification and intervention (particularly for children and adolescents) to mitigate long-term consequences ([Patel et al., 2018](#); [World Health Organization, 2022c](#)). Nevertheless, large treatment gaps persist: only a minority of people with a major depressive disorder receive minimally adequate care, even in high-income countries ([Thornicroft et al., 2017](#)). On the other hand, social media has emerged as a means to help monitor mental health. Many individuals disclose emotions and struggles on platforms such as Reddit and X, which could potentially contain linguistic markers associated with depression, self-harm, suicide ideation, and related conditions ([Chancellor & De Choudhury, 2020](#); [Coppersmith, Dredze, & Harman, 2014](#); [De Choudhury, Gamon, Counts, & Horvitz, 2013](#); [Husseini Orabi, Buddhitha, Husseini Orabi, & Inkpen, 2018](#); [Skaik & Inkpen, 2020](#); [Yates, Cohan, & Goharian, 2017](#)).

The availability of large-scale user-generated text has motivated research into automated detection of mental health risks from social media ([Skaik & Inkpen, 2020](#)). If models can flag at-risk users

early, they may enable more timely referral to support or intervention (Buddhitha & Inkpen, 2023). Benchmarks such as the CLEF eRisk datasets allow researchers to evaluate their developed methods for both correctness and earliness in sequential post-by-post settings (Losada, Crestani, & Parapar, 2019, 2020; Parapar, Martín-Rodilla, Losada, & Crestani, 2021, 2022, 2023, 2024). However, building reliable systems in this space remains challenging. Language used in social media is informal and full of slang, abbreviations, misspellings, emojis, etc., requiring robust Natural Language Processing (NLP) (Amini, Farahnak, & Kosseim, 2019; Baldwin et al., 2015; Eisenstein, 2013). Signals of mental illness can also be rare and limited to a small subset of posts within a user’s history (Yates et al., 2017), and any real-world use raises concerns around privacy, consent, and potential misuse (Chancellor, Birnbaum, Caine, Silenzio, & De Choudhury, 2019; Chancellor & De Choudhury, 2020).

A particularly critical challenge is *explainability*. Clinicians are unlikely to trust black-box predictions in high-stakes settings without human-understandable justification, and clinical adoption literature consistently emphasizes the need for case-level, context-aware explanations (Amann, Blasimme, Vayena, Frey, & Madai, 2020; Tonekaboni, Joshi, McCradden, & Goldenberg, 2019). Beyond improving accuracy, explainability can help identify model failures and biases before deployment. In health and other safety-critical domains, interpretable-by-design approaches are preferable to post-hoc rationalizations of opaque models (Rudin, 2019). Consequently, our goal is not only to achieve strong detection performance, but also to provide transparent outputs (e.g., highlighting the specific posts that most influenced the decision), ideally those that align with clinical expectations (Tonekaboni et al., 2019).

1.2 Problem Statement

In this thesis, we address the problem of detecting mental health risks in social media posts using deep learning, with an emphasis on model interpretability. The core task can be formulated as a text classification problem: given a sequence of a user’s social media posts, the model must determine whether the user is at risk of a certain mental health condition (such as depression, anxiety, eating disorder, etc.). We are particularly interested in *early detection*, meaning the model should make

accurate predictions as early as possible in the post sequence, which is crucial for timely intervention. This entails tackling an *incremental* classification setting where data appears over time and decisions must be made with partial information.

Several challenges define the scope of this problem. First, the data is inherently noisy and user-generated. Language on social platforms includes slang, emojis, misspellings, and contains diverse styles, requiring robust NLP approaches [Amini et al. \(2019\)](#); [Baldwin et al. \(2015\)](#); [Eisenstein \(2013\)](#). Secondly, datasets for mental health evaluation are often small relative to typical deep learning needs, due to privacy and annotation constraints, which can lead to overfitting ([Chancellor et al., 2019](#); [Chancellor & De Choudhury, 2020](#); [Skaik & Inkpen, 2020](#)). These two challenges motivate the use of transfer learning and domain adaptation, which have shown clear benefits for social media and mental health texts ([Aragon, Lopez Monroy, Gonzalez, Losada, & Montes, 2023](#); [Gururangan et al., 2020](#)). Thirdly, there are significant class imbalances: only a small fraction of users on a platform or in a benchmark may exhibit the target condition, complicating discrimination between true signals and background noise ([Losada & Crestani, 2016](#); [Yates et al., 2017](#)). Last but not least, the *black-box* nature of deep neural networks is not aligned with the requirement for transparency in healthcare, as high-performing models that cannot explain their reasoning are hard to trust or adopt in clinical or safety-critical contexts ([Amann et al., 2020](#); [Rudin, 2019](#); [Tonekaboni et al., 2019](#)).

To summarize, this thesis addresses the following problem: *How can we develop a system for early detection of mental health issues in social media that is both accurate and explainable?* We approach this question by exploring the use of attention-based neural architectures which can identify the most informative posts as explanations, and strategies to validate that the provided explanations are faithful to the model’s decision process.

1.3 Research Objectives

Given the problem explained in Section 1.2, our objective is to design and evaluate an *explainable* early detection framework for mental health risks on social media. Beyond strong predictive performance, we aim to provide transparency on the model’s decision-making process.

We leverage attention mechanisms ([Bahdanau, Cho, & Bengio, 2015](#)) to (i) prioritize informative

posts and (ii) provide human-understandable rationales; and we *quantitatively* test whether such rationales are faithful to the model’s decision logic (Jacovi & Goldberg, 2020; Moradi, Kambhatla, & Sarkar, 2021).

We tackle this objective through three research questions (RQs) which align with the thesis’s goal:

RQ1 (Chapter 3). Can we build a model that performs strongly on early risk detection of mental health issues on social media?

RQ2 (Chapter 4). Can attention weights provide faithful explanations in text classification tasks, and can their faithfulness be quantified reliably?

RQ3 (Chapter 5). Can we develop an approach that simultaneously achieves competitive early risk detection and produces explanations with demonstrated faithfulness?

1.4 Overall Methodology

Our methodology follows the three research questions above and is structured as follows:

Step 1 - Establish a strong early-detection approach (RQ1 - Chapter 3). Guided by the literature (Chapter 2), we implement a hierarchical attention model for user-level early risk detection, i.e., using Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) with contextual and non-contextual word embeddings, on the eRisk 2019 dataset (Losada et al., 2019) that emphasizes earliness (latency-weighted F1). This step identifies effective modeling choices for the sequential, post-by-post setting which eRisk focuses on.

Step 2 - Quantify attention faithfulness (RQ2 - Chapter 4). We then investigate whether attention weights can provide *faithful* explanations. Building on adversarial attention analyses, we introduce a single-run adversarial setup that yields a quantitative faithfulness score for attention, called ATTEFA. This novel metric lets us measure, across synthetic and real datasets, whether attention highlights decision-relevant input elements (tokens/posts) or can be perturbed without affecting predictions.

Step 3 - Propose a user-aware attention architecture (RQ3 - Chapter 5). Finally, we integrate a disorder-aware transformer encoder, DISORBERT (Aragon et al., 2023), with hierarchical, user-level attention to produce both a classification and an explanation in the form of highlighted posts. We perform ablation studies and compare against eRisk shared tasks’ participating models, assessing both early-detection performance and explanation faithfulness (via ATTEFA) and analyzing which posts drive decisions.

1.5 Contributions

This thesis makes the following contributions, organized by the research questions presented in Section 1.3:

- (1) We introduce a hierarchical attention model for user-level early risk detection on social media timelines within the eRisk evaluation framework (Losada & Crestani, 2016; Parapar et al., 2024). Our approach achieved first place in the eRisk 2019 shared task (Losada et al., 2019) on early detection of anorexia. Chapter 3 describes the architectures and training setup, the decision protocol and metrics, ablations isolating effective modeling choices, and an analysis of attention as post-level evidence. This chapter is based on our published book chapter in *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the eRisk Project* (Amini et al., 2022).
- (2) We introduce a *single-run adversarial* setup to compute a novel metric called ATTEFA, enabling *quantitative* assessment of whether attention weights are faithful in the decision process. We validate ATTEFA on controlled synthetic data and real text datasets, revealing when attention can/cannot be trusted as an explanation. Chapter 4, based on our published paper in the proceedings of the BlackBoxNLP 2022 workshop (Amini & Kosseim, 2022), focuses on this method and the empirical findings.
- (3) We modernize and strengthen the model from Chapter 3 and propose a hierarchical architecture that integrates the domain-adapted transformer encoder model, DISORBERT (Aragon et al., 2023), with a user-level attention mechanism. This design enables the model to emphasize

the social media posts that are most indicative of mental health risk. The system achieves competitive early-detection performance while producing post-level explanations, whose faithfulness we quantify using ATTEFA (introduced in Chapter 4). Our ablation studies further show that only a small subset of posts typically drives the model toward a positive (i.e., at-risk) prediction. Chapter 5 presents the architecture, experiments, ablations, and explanation analyses. This contribution was published in the proceedings of the RANLP 2025 conference (Amini & Kosseim, 2025).

- (4) Across chapters, we propose practical recommendations on when attention is (and is not) reliable as an explanatory signal, and how to combine attention with quantitative faithfulness tests (e.g., ATTEFA) when explainability is a requirement.

1.6 Thesis Organization

Organized around the three research questions in Section 1.3, the thesis is structured as follows:

- **Chapter 2** presents background and related work on detecting mental-health risks from social media and on explainability methods in NLP. It motivates the design choices used in the rest of the thesis.
- **Chapter 3 (RQ1)** presents our early-detection system for anorexia within the eRisk framework. We report the official eRisk 2019 results, together with ablation studies and an initial analysis of attention as post-level evidence.
- **Chapter 4 (RQ2)** presents our single-run adversarial setup to compute our novel ATTEFA metric, which quantitatively assesses attention faithfulness. We evaluate the method on synthetic and real datasets and discuss when attention is, and is not, faithful.
- **Chapter 5 (RQ3)** presents our user-aware attention framework built on DISORBERT. We report results on four eRisk tasks (depression, anorexia, self-harm, and pathological gambling) and analyze the faithfulness and robustness of the produced explanations.
- **Chapter 6** summarizes the contributions, limitations, and directions for future work.

Chapter 2

Literature Review

2.1 Introduction

Mental health detection from social media has emerged as a promising direction at the intersection of Natural Language Processing (NLP) and clinical psychology. The core idea is to analyze user-generated text to identify early markers of conditions such as depression, anxiety, and suicide risk, so that timely support can be offered when needed. Unlike clinical visits, social media lets us study how people use language over time in real life. However, since the posts are informal and inconsistent, more robust NLP models are necessary and tests to handle noisy and informal texts are needed ([Chancellor & De Choudhury, 2020](#); [Coppersmith et al., 2014](#); [Parapar et al., 2024](#)).

Early systems mostly used hand-crafted features, such as word lists and writing style measures, to train feature-based classifiers. These works showed that such systems were feasible, but they depended heavily on manual features and did not generalize well ([Coppersmith et al., 2014](#); [DeVault et al., 2013](#); [Pestian, Nasrallah, Matykiewicz, Bennett, & Leenaars, 2010](#); [Shen & Rudzicz, 2017](#)). More recently, deep learning became central to the field, beginning with convolutional and recurrent architectures ([Husseini Orabi et al., 2018](#); [Yates et al., 2017](#)). From there, two main directions emerged: 1) hierarchical and user-level models designed to capture the structure of multi-post timelines ([Maupomé, Queudot, & Meurs, 2019](#); [Mohammadi, Amini, & Kosseim, 2019a](#); [Mohammadi et al., 2019b](#)), and 2) transformer-based models pre-trained at scale to embed richer language representations ([Aragon et al., 2023](#)). More recently, combinations of these approaches have also

been explored, such as integrating hierarchical attention with domain-adapted transformers (Amini & Kosseim, 2025). Alongside these advances came growing attention to ethical, privacy, and deployment considerations (Chancellor et al., 2019) and to model *explainability*, which is central for clinician and moderator trust (Rudin, 2019; Tonekaboni et al., 2019).

This chapter surveys the literature most relevant to this thesis. Section 2.2 outlines background on mental-health NLP and shared tasks that have shaped the field. Section 2.3 reviews deep learning methods for detection, including early neural architectures, hierarchical/user-level models, pre-trained transformers, and a discussion of Large Language Models (LLMs) and why they are not adopted as core models in this thesis. Section 2.4 covers explainability methods for NLP, with an emphasis on attention-based explanations and their validation, and on explainability in mental-health applications. We conclude with a summary that motivates the approach developed in later chapters.

2.2 Background

NLP for mental health uses language as a way to understand people’s well-being. A wide range of platforms (e.g., Reddit and X) and conditions (e.g., depression, anxiety, PTSD, and self-harm risk) have been explored, demonstrating that patterns in language correlate with mental health states (Chancellor & De Choudhury, 2020; Montejo-Ráez, Molina-González, Jiménez-Zafra, Ángel García-Cumbreras, & García-López, 2024). Early pipeline designs used curated lexicons, sentiment or affective features, and stylistic measures to train linear models (Coppersmith et al., 2014; DeVault et al., 2013; Pestian et al., 2010; Shen & Rudzicz, 2017). Although these studies showed such a task was feasible, they often failed to transfer well between communities and missed subtle signals in short, highly informal posts.

Community benchmarks catalyzed progress. The CLPsych workshop series introduced shared tasks on detecting mental health signals from social media, creating common datasets and evaluation settings (Chim et al., 2024; Coppersmith, Dredze, Harman, Hollingshead, & Mitchell, 2015; Lynn et al., 2018; Milne, Pink, Hachey, & Calvo, 2016; Tsakalidis et al., 2022; Tseriotou et al., 2025; Zirikly, Resnik, Uzuner, & Hollingshead, 2019). A few years later, the CLEF eRisk Lab introduced *early risk detection* as part of the eRisk shared tasks (Losada, Crestani, & Parapar, 2017, 2018;

[Losada et al., 2019, 2020](#); [Parapar et al., 2021, 2022, 2023, 2024](#)): systems must process a user’s timeline chronologically and trigger an alert as soon as sufficient evidence for the mental health issue accumulates. This reframed detection as a sequential decision problem, with metrics that balance timeliness and correctness. These tasks also highlighted practical constraints: small, noisy labeled datasets; demographic and domain shift; and the need to provide actionable, transparent outputs that align with clinical workflows ([Chancellor & De Choudhury, 2020](#); [Tonekaboni et al., 2019](#)).

2.3 Deep Learning for Mental Health Detection

By the mid-2010s, deep learning started to replace NLP systems relying on hand-crafted features ([Amini et al., 2019](#)). Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) learned task-relevant representations directly from text, improving robustness to slang and non-standard phrasing. For instance, [Yates et al. \(2017\)](#) modeled user posts with CNNs and aggregated signals for user-level risk. [Husseini Orabi et al. \(2018\)](#) compared CNNs and Long Short-Term Memory (LSTM) networks and reported consistent gains over a linear Support Vector Machine (SVM) baseline. These models reduced reliance on manual features, captured local patterns and some temporal dependencies, and established a new baseline for subsequent research.

User timelines are naturally hierarchical: many words compose posts and many posts compose a user model. Earlier systems either flattened all posts into a single document, losing timeline structure, or made predictions per-post and pooled predictions into a user label, which is noisy because labels are user-level and many posts are uninformative to mental health prediction. Hierarchical encoders (e.g., word $\rightarrow$ post $\rightarrow$ user) and explicit timeline models address this drawback by (i) producing post-level embeddings and (ii) aggregating them with sequence encoders or attention to produce user-level predictions. Inspired by Hierarchical Attention Networks (HANs) ([Z. Yang et al., 2016](#)), this line of work improved both performance and usability by surfacing salient posts for inspection ([Maupomé et al., 2019](#); [Mohammadi et al., 2019a, 2019b](#)).

Transformer architectures ([Vaswani et al., 2017](#)) and large-scale pre-training transformed NLP. Fine-tuning pretrained transformer models, such as BERT ([Devlin, Chang, Lee, & Toutanova, 2019](#)), quickly became standard for mental-health text classification, improving accuracy and recall in

data-constrained settings (Murarka, Radhakrishnan, & Ravichandran, 2021). Domain-adaptive pre-training (Gururangan et al., 2020) further mitigated mismatch between general corpora and social media text. In mental health contexts, domain adaptation and task-specific designs yielded sizable gains; for example, adapting BERT to social-media mental-health content improved early detection across depression and self-harm tasks (Aragon et al., 2023).

The promise of both hierarchical attention architectures, which can highlight the most informative posts in a user’s timeline, and pre-trained domain-adapted language models, which provide robust contextual representations of social media and mental health language, motivated us to pursue a hybrid approach combining the two. This synergy allows the model to benefit from powerful encoders that capture nuanced linguistic cues, while retaining the ability to identify post-level explanations through attention. In Chapter 5, we present such a user-aware framework that integrates hierarchical attention with a domain-adapted transformer, demonstrating how this hybrid design yields both competitive predictive performance and faithful explanations of model decisions.

LLMs demonstrate remarkable few-shot capabilities via prompting (Brown et al., 2020). Building on this, several recent studies have applied LLMs to mental health detection on social media: for example, Y. Wang, Inkpen, and Kirinde Gamaarachchige (2024) used LLaMA-2 (Touvron et al., 2023), SUS-Chat¹ and Neural-chat² for explainable depression classification aligned with Beck’s Depression Inventory (BDI) symptoms (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), and Bauer et al. (2024) leveraged GPT-4 (Brown et al., 2020) for suicidality analysis on Reddit posts. However, for *early, explainable* risk detection on social media, there are compelling reasons not to adopt LLMs as core models in this thesis:

- **Domain adaptation vs. prompting.** Specialized domain adaptation and post-level aggregation (Section 2.3) reliably outperform generic zero-/few-shot prompting in narrow tasks with limited labels, especially under distribution shift (Aragon et al., 2023; Gururangan et al., 2020).
- **Explainability requirements.** Clinical stakeholders require transparent, verifiable rationales (Rudin, 2019; Tonekaboni et al., 2019). While LLMs can generate plausible textual “explanations,” these are post-hoc narratives that may not be faithful to the model’s internal decision

¹<https://huggingface.co/SUSTech/SUS-Chat-34B>

²<https://huggingface.co/Intel/neural-chat-7b-v3-1>

basis. Our approach emphasizes mechanisms (e.g., attention over posts, symptom-aligned outputs) whose faithfulness can be stress-tested.

- **Fine-tuning and data leakage risks.** Fine-tuning generative LLMs on raw user timelines in this domain carries concrete privacy risks: LLMs can sometimes “memorize” pieces of the text they were trained on and later repeat them word-for-word (Carlini et al., 2020). They are also at risk of *membership inference attacks* (Shokri, Stronati, Song, & Shmatikov, 2017), where an adversary can test or prompt the model to determine whether a specific record was in the training set or even recover memorized examples (Ishihara, 2023). Because mental-health content is public-but-sensitive and governed by ethical/legal constraints, curating a dataset suitable for fine-tuning LLMs requires strict de-identification, consent/licensing checks, and auditing, making it significantly more challenging than fine-tuning compact task-specific encoders (Chancellor et al., 2019).
- **Cost and latency.** LLMs inference can be financially costly and environmentally intensive, requiring large amounts of computation and energy. For large-scale, near-real-time screening, compact models fine-tuned and deployed on controlled infrastructure offer predictable latency and cost.

In short, given our goals (early detection, rigorous evaluation, and faithful, inspectable explanations) domain-adapted transformers with user-level aggregation provide a more appropriate balance of performance, governance, and transparency than current LLM workflows.

2.4 Explainability in NLP Models

Explainability methods in NLP are often organized along two axes: 1) *intrinsic* vs. *post-hoc*, and 2) *local* vs. *global* (Koh & Liang, 2017; Lundberg & Lee, 2017; Ribeiro, Singh, & Guestrin, 2016; Rudin, 2019).

Intrinsic (interpretable-by-design). The model itself is understandable without extra tools. Examples include small decision trees, sparse linear/logistic models with human-meaningful features,

additive models, and rule lists. In NLP, bag-of-words or lexicon-based linear models are often considered intrinsic because their coefficients or rules can be read directly as explanations (Rudin, 2019).

Post-hoc (explanations after training). Explanations are produced for an already-trained, often black-box, model without changing it. Common approaches include feature attribution, example-based explanations, and attention visualizations. Post-hoc methods are prevalent because high-performing neural models are rarely interpretable by design (Rudin, 2019).

Local (per-input) explanations. These explain a single prediction, such as why a model flagged one Reddit post as “at risk.” Several popular techniques fall in this category:

- *LIME* (Ribeiro et al., 2016) builds a *local surrogate model* around the instance of interest. This means the model perturbs the input (e.g., removes or masks words), observes how the black-box model changes its prediction, and then fits a simple interpretable model (like a sparse linear regression) that approximates the black box only in the neighborhood of that instance.
- *SHAP* (Lundberg & Lee, 2017) assigns each input feature (e.g., a word) a Shapley value (Shapley, 1953), which comes from cooperative game theory. Intuitively, the Shapley value measures how much that feature contributes to the prediction by averaging how much the feature helps the model’s prediction when it is included alongside different groups of other features.
- *Influence functions* (Koh & Liang, 2017) estimate how much each training example influenced the prediction for a given input. For instance, they can identify which training posts most increased the likelihood that the model labeled a user as depressed. This provides a case-based explanation rather than a feature-based one.

Global (model-level) explanations. Such explanations summarize how the model behaves across many inputs. They can answer questions such as which features matter most overall, or what kinds of posts are generally influential. Common approaches include training a global surrogate model (e.g.,

a decision tree that approximates the black box across the dataset), aggregating local attributions into summary plots (as is often done with SHAP), or running dataset-wide ablations to measure the effect of removing groups of features (Lundberg & Lee, 2017).

It is worth noting that some methods, like SHAP, can serve both as local and global explanations: the same Shapley values can be inspected at the instance level (local) or averaged across the dataset to see global trends.

Because our deep learning models are not intrinsically interpretable, we will primarily use *post-hoc, local* explanations at the user/post level and then aggregate them for global insight.

2.4.1 Aspects of Explanation

In the literature, four complementary aspects of explanation are often distinguished: *transparency*, *plausibility*, *sufficiency*, and *faithfulness* (Amini & Kosseim, 2022; Lipton, 2016; Rudin, 2019; Wiegrefe & Pinter, 2019).

Transparency. Transparency refers to how much of the model’s internal workings are directly understandable to humans, e.g., highlighting which posts or tokens the model attends to (Lipton, 2016). Attention weights, for instance, can provide partial transparency by revealing what the model focuses on.

Plausibility. Plausibility means that explanations appear reasonable and convincing to human observers, even if they are not causally faithful to the model’s decision process (Rudin, 2019; Wiegrefe & Pinter, 2019). Human evaluations are typically used to assess plausibility.

Sufficiency. An explanation is sufficient if the highlighted parts of the input alone are enough for the model to make the same prediction. This is often tested by masking all but the rationale and checking whether the prediction is preserved (DeYoung et al., 2020).

Faithfulness. Faithfulness requires that the explanation truly reflects the model’s decision basis, i.e., altering or removing the highlighted features should reliably change the model’s output (Amini & Kosseim, 2022; Jain & Wallace, 2019; Serrano & Smith, 2019). Our novel ATTEFA metric (Amini & Kosseim, 2022) directly measures this property (see Chapter 4).

In this thesis, we focus on **transparency** (by surfacing the most informative posts for each user)

and especially on **faithfulness** (through systematic evaluation with ATTEFA). Full assessments of **plausibility** and **sufficiency** are left for future work.

2.4.2 Attention as a Means for Explanation

Attention mechanisms assign weights to input tokens or posts according to their contribution to the task and are frequently visualized as explanations (Bahdanau et al., 2015). Attention visualization is often appealing because it tends to highlight evidence that is comprehensible by humans (e.g., posts where a user openly expresses hopelessness). However, it is debated whether attention can truly be considered as an explanation. Jain and Wallace (2019) showed that alternate attention distributions can yield similar predictions, and that attention weights may misalign with other importance measures. Wiegrefe and Pinter (2019) argued that, under certain constraints, attention can still be informative. Other works demonstrated that attention maps can be adversarially manipulated without affecting accuracy, emphasizing the need for caution (Pruthi, Gupta, Dhingra, Neubig, & Lipton, 2020; Serrano & Smith, 2019).

In response, researchers proposed diagnostics and training strategies to assess or improve *faithfulness*. For example, TaSC (Chrysostomou & Aletras, 2021) evaluates how faithful attention weights are by testing the model’s sensitivity to controlled interventions: it deliberately alters the attention distribution (e.g., forcing the model to focus on tokens it originally ignored) and then checks whether the prediction changes accordingly. If the output remains largely the same despite these changes, it suggests that the original attention weights were not truly driving the model’s decision.

In Chapter 4, we will propose the ATTEFA metric (Amini & Kosseim, 2022), which quantifies attention faithfulness by asking a simple question: Can a model keep making the same predictions while being forced to attend elsewhere? We achieve this with a single-run adversarial setup that trains a companion model to match the base model’s outputs as closely as possible while pushing its attention distribution far away; if the outputs still align, attention is likely not driving decisions (low ATTEFA), whereas if the outputs cannot be preserved under large attention shifts, attention is more faithful (high ATTEFA). We validate the metric on synthetic and real datasets and analyze how encoder contextualization and the sparsity of informative tokens affect faithfulness.

2.4.3 Explainability in Mental Health NLP

Explainability is critical for the proper adoption of methods in clinical and safety-critical settings. Beyond highlighting tokens, mental-health systems often align outputs with clinical concepts. For example, symptom-oriented models predict BDI categories (Beck et al., 1961), such as sadness, insomnia, as intermediate signals for severity estimation, yielding human-interpretable decompositions (Inkpen, Skaik, Buddhitha, Angelov, & Fredenburgh, 2021; Skaik & Inkpen, 2022). Other works map evidence to psychologically motivated constructs such as cognitive distortions, providing theory-grounded rationales (B. Wang, Zhao, Lu, & Qin, 2023). These strategies help moderators and clinicians triage alerts and prioritize interventions, while maintaining trust (Chancellor & De Choudhury, 2020; Tonekaboni et al., 2019).

In summary, effective explainability in this domain entails 1) providing concise, user-level evidence (e.g., a few most relevant posts), 2) validating explanation faithfulness (rather than assuming it), and 3) when possible, aligning outputs with clinically meaningful constructs. The approach in this thesis adopts the first two, while providing guidelines on how to tackle the third item (in Chapter 6).

2.5 Conclusion

In this chapter, we reviewed prior work in two main areas: 1) deep learning methods for detecting mental-health risks on social media, and 2) explainability techniques for NLP models, with particular attention to mental-health applications. We traced the field’s progression from feature-based baselines, to RNNs and CNNs, to hierarchical user-level models, and finally to transformer-based systems adapted to social media and aligned with clinical constructs. We also explained why this thesis does not adopt LLMs as core models, given the goals of early detection and the need for explanations that are both faithful and auditable.

On the explainability side, we summarized feature attribution methods, example-based approaches, attention-based rationales and their critiques, and clinically grounded strategies. These perspectives inform the design choices for the remainder of the thesis: using compact, domain-adapted transformers with user-level aggregation, providing transparent evidence for predictions, and evaluating explanations through systematic tests of faithfulness.

Chapter 3

Early Detection of Anorexia in Social Media

In this chapter, we aim to address the first research question presented in Section 1.3:

Can we build a model that performs strongly on early risk detection of mental health issues on social media?

An earlier version of this chapter appeared as an invited extended chapter in the book *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the eRisk Project* (Amini et al., 2022). The book chapter consolidated and extended our previously published work on our participation in the eRisk 2019 shared task (Mohammadi et al., 2019b), and our subsequent publication at the NLDB 2020 conference (Amini & Kosseim, 2020).

This chapter overviews our participation to the first sub-task of eRisk 2019 shared task, which focused on the task of early detection of anorexia in Reddit posts. Our approach utilized several attention-based neural sub-models to extract features and predict class probabilities, which were later used as input features to a Support Vector Machine (SVM) making the final classification. Our submission, named *CLaC* achieved F1 and latency-weighted F1 scores of 0.7073 and 0.6908 respectively, allowing it to rank first in terms of these metrics, and achieved competitive results based on other evaluation metrics. This chapter also extends our participation to eRisk 2019, and includes additional experiments that were performed on the model in order to evaluate it in terms of

explainability.

3.1 Introduction

In the last decade, the use of social media to express personal thoughts, emotions, and ideas has become more and more prevalent. The analysis of online data can be useful for many purposes, such as business and marketing, political planning, prediction of stock market ([Giatsoglou et al., 2017](#)), as well as enhancing awareness of emergencies ([Yin et al., 2015](#)). Another noteworthy line of research has focused on the detection of toxicity, hate speech, aggression and cyber bullying on online platforms; an effort that could facilitate timely interventions in violent situations ([Davidson, Warmlesley, Macy, & Weber, 2017](#); [Thompson, Leung, Blair, & Taboada, 2017](#)).

In healthcare applications, online posts have been used for detecting disease outbreaks ([Ofoghi, Mann, & Verspoor, 2016](#)), finding smoking patterns ([Struik & Baskerville, 2014](#)), and the identification of adverse drug reactions ([C. C. Yang, Jiang, Yang, & Tang, 2012](#)). Another useful application is the automatic detection of mental health issues, a relatively recent field which has attracted the attention of many researchers in Natural Language Processing (NLP). Corpora from X, Facebook, blogs and online forums, and Reddit are used as resources to detect various mental health problems, such as anxiety, depression, suicide ideation, and eating disorders ([Calvo, Milne, Hussain, & Christensen, 2017](#)).

Although automatically monitoring online forums to detect cases of mental health issues is beneficial, the elapsed time between the first signs of a mental issue and the actual detection of a potential victim can play a crucial role. Earlier detection of a harmful behavior can help moderators better handle the situation. However, to the best of our knowledge, not much research has specifically addressed the task of early detection of mental health issues.

The eRisk shared task ([Losada et al., 2017](#)) was created with the goal of addressing issues related to early risk detection of mental health problems. According to [Losada et al. \(2017\)](#), early detection can be useful in many applications from the identification of potential sexual offenders to the detection of victims of suicidal tendencies, making intervention possible before it is too late. [Losada et al. \(2017\)](#) argues that while current risk assessment approaches often aim at detecting

harmful behavior after the fact, it is very important to consider the timing of risk detection and to minimize the time between the observation of the first evidence of destructive behavior and triggering an alarm. To that end, the organizers of the eRisk shared task have encouraged the development of approaches which model the process rather than the outcome, as well as developing reliable evaluation metrics and test collections tailored to early risk detection.

The aim of this chapter is to overview our participation to the eRisk 2019 shared task (Losada et al., 2019) and study the approach designed for the early detection of anorexia, in terms of explainability. In that sense, this work can be considered as an extension to Mohammadi et al. (2019b).

3.2 Related Work

Many researchers have used corpora from X, Facebook, Reddit, blogs and online forums as resources to experiment with classification tasks pertaining to mental health issues (Calvo et al., 2017).

Pestian et al. (2010) experimented with different machine learning methods for suicide note classification. The features used in the study included words, part of speech tags, readability scores, and emotions. The best accuracy of 74% was achieved by a logistic regression model.

DeVault et al. (2013) studied the symptoms of psychological distress in dialogues with a virtual agent. The use of a Naïve Bayes classifier for the detection of Post-Traumatic Stress Disorder (PTSD) and distress yielded a 20% improvement over the baseline accuracy of 53.5%, and showed that the automatic assessment of psychological distress is indeed possible.

More recently, Jackson et al. (2017) used clinical texts obtained through the Clinical Record Interactive Search¹ to extract symptoms of severe mental illness. The authors made use of TextHunter (Ball, Patel, Hayes, Dobson, & Stewart, 2014) (a NLP information extraction tool) and an SVM classifier, and were able to classify 38 symptoms with an F1-score of 85%.

Shen and Rudzicz (2017) used different feature sets including Word2Vec embedding (Mikolov, Chen, Corrado, & Dean, 2013), Latent Dirichlet Allocation (LDA) topic modelling (Blei, Ng, &

¹<https://crisnetwork.co>

[Jordan, 2003](#)), lexico-syntactic features, and n-grams (unigrams and bigrams) to detect anxiety in Reddit posts. Initially, the authors compared the results achieved by an SVM and a 2-layer neural network. Though both classifiers performed well, the SVM yielded marginally better results. However, they achieved their best result of 98% accuracy using the neural network with n-gram probabilities and word embeddings combined with Linguistic Inquiry and Word Count (LIWC) features.

[Coppersmith et al. \(2014\)](#) explored the automatic detection of PTSD, depression, bipolar disorder, and seasonal affective disorder (SAD) in X data, using LIWC features and character and word n-grams, and found the latter resulting in superior performance.

[Benton, Mitchell, and Hovy \(2017\)](#) used multi-task learning to predict suicide risk and a variety of mental health conditions from X data, including anxiety, depression, PTSD, and schizophrenia. It was found that a multi-task framework can be effectively used in cases with limited data.

Apart from individual efforts, shared tasks (e.g., [Coppersmith et al., 2015](#); [Milne et al., 2016](#); [Lynn et al., 2018](#); [Zirikly et al., 2019](#)) have also been organized to encourage the development of common benchmarks (datasets and metrics) and the comparison of approaches for the detection of distress in online textual data.

All of the work described above used a classic classification approach that does not measure how early the detection is performed. On the other hand, the eRisk shared tasks ([Losada et al., 2017, 2018, 2019, 2020](#)) focus on the early detection of mental health issues. In the first edition of eRisk ([Losada et al., 2017](#)), the data set used was a collection of social media posts and comments from depressed and non-depressed authors, recorded chronologically. As evaluation metric, Early Risk Detection Error (ERDE; see Section 3.4) was used, an error measure which assigns a penalty to late decisions and rewards early ones ([Losada & Crestani, 2016](#)). As it was the first edition of this shared task, many teams focused on making accurate rather than early decisions, with the highest F1-score being 64% and the lowest $ERDE_{50}$ score being 9.68% ([Losada et al., 2017](#)).

The second eRisk shared task ([Losada et al., 2018](#)) included two sub-tasks: Early risk detection of depression and early risk detection of anorexia. Like the year before, the ERDE evaluation metric was used as the main metric alongside F1, precision, and recall ([Losada et al., 2018](#)). The best performing systems, in both sub-tasks, were designed by [Trotzek, Koitka, and Friedrich \(2018\)](#).

Their team experimented with different variations of bag of words features and a Convolutional Neural Network (CNN) (LeCun, Haffner, Bottou, & Bengio, 1999) as well as ensemble models. In the depression sub-task, their system achieved an F1-score of 64% and an $ERDE_{50}$ of 6.44%. In the anorexia sub-task, they achieved an F1-score of 85% and an $ERDE_{50}$ of 5.96%.

In eRisk 2019, the release of the data in the testing phase was changed from chunk-based (i.e., the method used in the previous eRisk shared tasks) to one-by-one. In this process, a decision by the models could be emitted at any point, but once a user was detected as positive by a model, it was not possible to revert back to negative after having more observations. The new evaluation metric *latency-weighted F1* was introduced by the organizers as the official metric, which was essentially an F1 score in time series, with an incremental penalty factor over time. The main reason for introducing a new evaluation metric was the fact that, ERDE was not capable of properly handling the new testing scheme, as the penalty factor in ERDE increased very quickly in one-by-one release of test data. The first sub-task of eRisk 2019 was focused on the early detection of anorexia, where our submission (Mohammadi et al., 2019b) achieved the highest latency-weighted F1 score of 0.6908 and significantly outperformed the second team with a margin of around 6%.

In this chapter, we will revisit our participation to eRisk 2019 shared task, and will summarize the designed approach and results. We will also study our model in terms of its capability to provide explainability.

3.3 Task and Dataset

For this work, we used the dataset from the first sub-task of eRisk 2019 (Losada et al., 2019), which was focused on the early detection of anorexia. The dataset was a collection of Reddit users labelled as anorexic or non-anorexic (Losada & Crestani, 2016), along with a collection of their Reddit posts, recorded chronologically.

For the training phase, the data from the previous year (eRisk 2018 task 2), including both training and test sets, was made available. For the testing phase, posts were released on an item-by-item basis in chronological order for a new collection of Reddit users. The goal was to detect users suffering from anorexia, having observed as few posts from them as possible. As a result, in addition to

precision, recall, and F1-score, two other metrics were used: ERDE measure which penalizes late decisions, and latency-weighted F1, a modified version of F1 score that takes into account the delay of the decision.

Table 3.1 shows the statistics of the datasets. As shown in the table, the datasets are highly imbalanced, with about 90% of the users not suffering from anorexia.

Dataset	Source	Positive	Negative	All
Training	Train 2018	20 (13%)	132 (87%)	152
Validation	Test 2018	41 (13%)	279 (87%)	320
Testing	—	73 (9%)	742 (91%)	815

Table 3.1: Distribution of user labels in the datasets. The 2018 datasets refer to the eRisk 2018 task 2 data.

3.4 Evaluation Metrics

3.4.1 Early Risk Detection Error (ERDE)

Similar to the previous eRisk shared tasks (Losada et al., 2017, 2018), ERDE was used as an error measure in the task of early detection of anorexia. This measure not only penalizes false positives and false negatives, but it also applies a delay penalty (which increases as the delay increases) to the true positive samples. The equation below shows how the $ERDE_o$ measure for each sample is calculated.

$$ERDE_o(d) = \begin{cases} c_{fp} & \text{if } d=\text{positive AND ground truth}=\text{negative (FP)} \\ c_{fn} & \text{if } d=\text{negative AND ground truth}=\text{positive (FN)} \\ l_{c_o}(k).c_{tp} & \text{if } d=\text{positive AND ground truth}=\text{positive (TP)} \\ 0 & \text{if } d=\text{negative AND ground truth}=\text{negative (TN)} \end{cases} \quad (1)$$

As Equation 1 shows, four different types of penalties will be applied to a classification, based on the actual label and the predicted one by a classifier:

- (1) For a false positive sample, the penalty c_{fp} is applied, which is equal to the proportion of positive samples in the test data.

- (2) In a false negative case, the penalty will be c_{fn} , which is equal to 1.
- (3) For a true positive sample, the penalty that is applied is equal to $lc_o(k) \cdot c_{tp}$. While c_{tp} is equal to 1, $lc_o(k)$ is calculated using Equation 2:

$$lc_o(k) = 1 - \frac{1}{1 - e^{k-o}} \quad (2)$$

where k is the number of evidences observed before the classifier detects a positive sample.

- (4) No penalty is applied to the system for true negative cases.

The final reported value for $ERDE_o$ is calculated by averaging over the calculated values of $ERDE_o$ for each sample. The $ERDE_5$ and $ERDE_{50}$ metrics are calculated by setting the parameter o in the $ERDE_o$ formula equal to 5 and 50, respectively.

3.4.2 Latency-weighted F1

The new evaluation metric *latency-weighted F1* was introduced by the organizers as the official metric, which was essentially an F1 score in time series, with an incremental penalty factor over time. The main reason for introducing a new evaluation metric was the fact that, ERDE was not capable of properly handling the new testing scheme, as the penalty factor in ERDE increased very quickly in the one-by-one release of test data.

Latency-weighted F1 is calculated by multiplying the original F1 by the metric *speed* (see Equation 3).

$$latency-weighted F1 = F1 \cdot speed \quad (3)$$

The parameter *speed* is calculated using Equation 4:

$$speed = (1 - \text{median}\{penalty(k_s) : s \in S, d_s = g_s = 1\}) \quad (4)$$

where d_s is the decision that the system has made for a specific data sample, while g_s is the sample's golden truth. The parameter $penalty(k_s)$ is the penalty for a true positive classification for

the sample, after observing k evidences. Equation 5 is used to calculate the parameter $penalty(k_s)$:

$$penalty(k_s) = -1 + \frac{2}{1 + e^{-p \cdot (k_s - 1)}} \quad (5)$$

where the parameter p is set in a way that it leads to the penalty being equal to 0.5 at the median number of observations for a sample.

3.5 System Overview

Figure 3.1 shows the architecture of the model that we used for the eRisk 2019 shared task. The full model includes 8 different neural sub-models, followed by a fusion component, which concatenates the neural features and predicted class probabilities from different sub-models, and forwards them to a final SVM classifier.

This section will provide a more detailed explanation of the different components of the model.

3.5.1 Sub-models

As shown in Figure 3.1, each sub-model includes an input layer that receives as input the posts by a user, and vectorizes its tokens using an embedding layer. The output of the input layer is then fed to a hidden layer, which is followed by a post-level attention/pooling layer that creates a representation of the post from its constituent tokens. The user-level attention layer is responsible for calculating the vector representation of the user, using her/his online posts. Finally, the output (classification) layer predicts the probability distribution of the positive and negative classes (i.e., anorexic versus non-anorexic).

Our main focus during the development of the sub-models was to include diversity of information sources, so that the final ensemble model can incorporate different points of views when performing the final classification.

Input Layer.

The inputs to the model are the online posts of each user. Each post is first tokenized and the tokens are sent to the word embedder, in order to be converted into dense vectors. As shown in

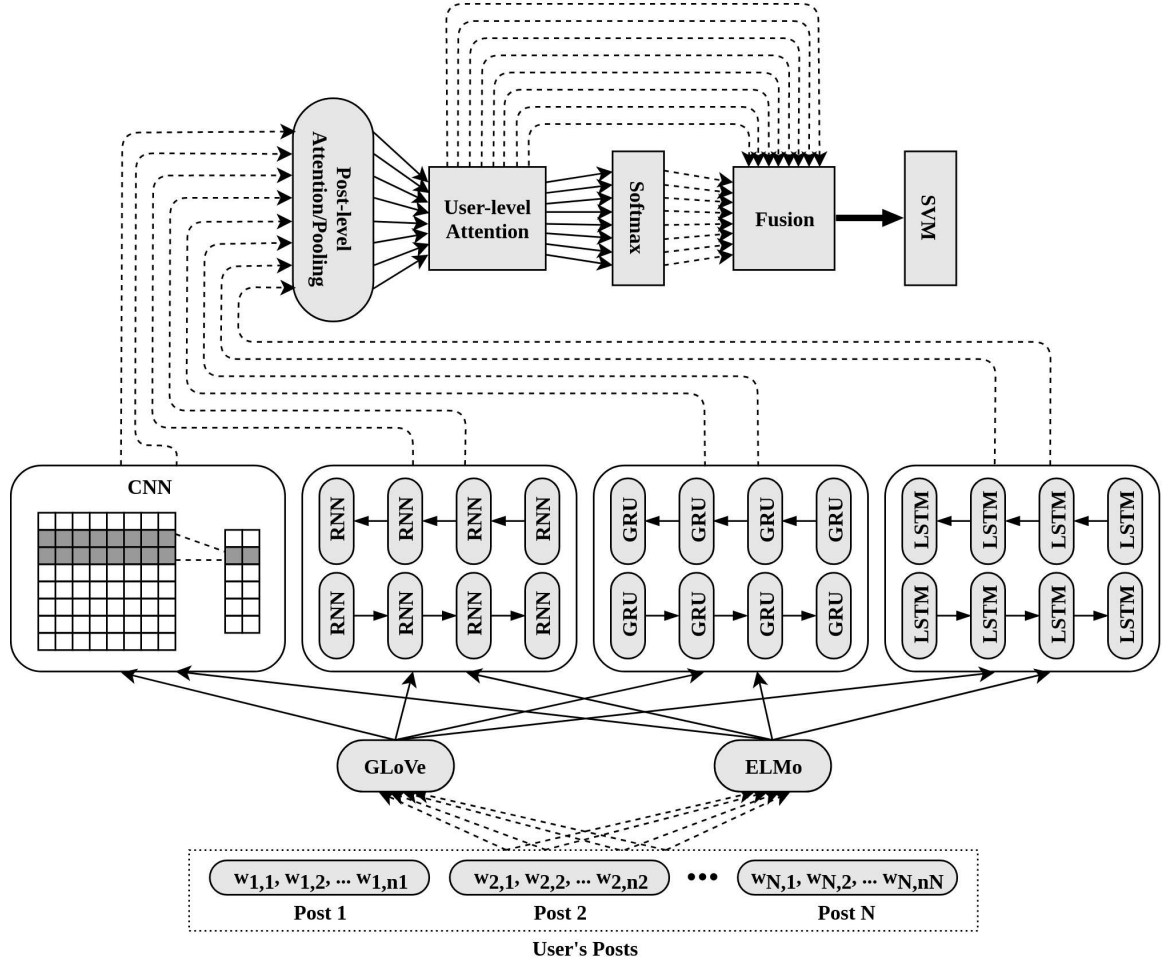


Figure 3.1: Architecture of the model. The number of arrows between components corresponds to the number of sub-models that move in that flow. The rounded-corner boxes represent the components that work at the post level, while the sharp-corner boxes are user-level components. The solid lines represent neural connections; while the dotted lines show the flow of data without the existence of a neural connection. The bold arrow between the Fusion and SVM corresponds to the flow of data that exists only in the final model.

Figure 3.1, these token vectors are then fed to the hidden layer.

Two different pretrained word embeddings were experimented with. The first word embedder was the 300d version of GloVe (Pennington, Socher, & Manning, 2014) that was pretrained on 840B tokens of web data from Common Crawl. The second word embedder was the original 1024d version of ELMo, which was pretrained on the 1 Billion Word Language Model Benchmark (Chelba et al., 2014). These two word embeddings were used in order to provide our ensemble model with sub-models that utilize both contextual (ELMo) and non-contextual (GloVe) word embedders in their input layer.

Hidden Layer

The hidden layer is responsible for processing the token vectors, generated by the input layer. As shown in Figure 3.1, we have experimented with four hidden architectures in our sub-models: a CNN (LeCun et al., 1999) that processes token n-grams separately, and a Bidirectional Vanilla Recurrent Neural Network (BiRNN), a Bidirectional Long Short-Term Memory (BiLSTM; Hochreiter and Schmidhuber, 1997) and a Bidirectional Gated Recurrent Unit (BiGRU; Cho et al., 2014), all of which process token vectors sequentially, from first to last and vice-versa, by taking into account the preceding and following tokens, respectively.

Post-level Attention/Pooling Layer

Following the work by Trotzek et al. (2018), for the sub-models that use CNN in the hidden layer, a max pooling is applied to the outputs of the hidden layer after being passed through a Concatenated Rectified Linear Unit (CReLU, i.e., ReLU applied on the concatenation of each vector and its negative).

In the models that use BiRNN, BiLSTM, or BiGRU in their hidden layer, an attention mechanism is responsible for computing the representation of a post (P) by weighted-averaging over the outputs of the hidden layer for each token in the post, where the weights assigned to each token is calculated automatically. The function used by the attention mechanism is shown in Equation 6:

$$P = \sum_{t=1}^n y_t \omega_t \quad (6)$$

where y_t represents the output of the recurrent hidden layer at time-step t , and ω_t is the weight assigned to the output in that time-step.

In our model, the attention mechanism uses an N -to-1 feed-forward layer (with the weights w , where N is equal to the size of the output vectors of the recurrent hidden layer) to map the output of the hidden layer at each time-step (e.g., y_t) to a scalar (e.g., ν_t):

$$\nu_t = y_t \times w \quad (7)$$

These scalars are then concatenated, and softmax is applied to the resulting vector. The resulting vector from the softmax will include the weights that are used by the attention mechanism:

$$\omega = \text{Softmax}([\nu_1, \nu_2, \nu_3, \dots, \nu_n]) \quad (8)$$

User-level Attention Mechanism

Knowing that the posts by a user do not contribute equally to evaluate her/his mental state ([Trotzek et al., 2018](#)), a user-level attention mechanism is used to have the system learn to automatically detect the contribution of each post to the final classification of the user.

The mechanism of the user-level attention is similar to the post-level attention mechanism, but it computes a vector representation of a user from the representation of her/his posts (resulting from the post-level attention/pooling).

Output (Classification) Layer

The final layer in the sub-models is a feed-forward fully-connected layer that maps the output of the user-level attention to a vector with size 2 (corresponding to the *negative* and *positive* classes). At the end of this layer, a softmax activation function gives as the output, the predicted probability distribution over the classes *negative* and *positive*.

3.5.2 Ensemble Model

As shown in Figure 3.1, the ensemble model is composed of several neural sub-models, a fusion component, and a final SVM classifier. The fusion component concatenates the outputs of the user-level attention units (which will subsequently be referred to as neural features), and the predicted probability distributions of the two classes, resulting from the softmax activation functions from all its constituent sub-models. The output of the fusion component is taken as the final representation of a user. This representation is finally fed to an SVM classifier to perform the ensemble classification.

3.6 Experimental Setup

This section describes our experiments with the above model for our participation to the eRisk 2019 shared task (Losada et al., 2019).

3.6.1 Sub-models Implementation

PyTorch (Paszke et al., 2019) was used to implement and train the sub-models. The Adam optimizer (Kingma & Ba, 2015) was used, and the learning rate was set to 5×10^{-4} . Cross-entropy was used as the loss function, and in order to handle the imbalanced distribution of the positive and negative classes in the training set (see Table 3.1), weights proportional to the inverse of the number of samples of each class were assigned to the classes. Due to limitations in terms of computational resources, mini-batches with a maximum size of 128 were used at the post level for each user and only the first 100 tokens of the posts were used². In order to minimize the amount of padding in the batches, posts with similar number of tokens were assigned to the same batch.

In order to fine-tune the other hyperparameters of the sub-models (including the number and size of convolutional filters, number of recurrent units, and number of training epochs), each sub-model was individually trained on the training set and optimized on the validation set (see Table 3.1), based on F1 score. The specifics of the 8 different sub-models are shown in Table 3.2. Since each sub-model is composed of a unique pair of hidden layer and word embedding type, they will later be

²This limit only truncated a small number of posts, as the average length was ~ 37.47 tokens in the eRisk 2018 task 2 data.

referred to as $\langle hidden-type \rangle - \langle embedding-type \rangle$ (see the second column of Table 3.2).

#	Name	Hyperparameters
1	<i>CNN-GloVe</i>	100 bigram convolution filters, trained for 10 epochs
2	<i>CNN-ELMo</i>	200 unigram filters and 50 bigram convolution filters, trained for 6 epochs
3	<i>BiRNN-GloVe</i>	one layer of 64 vanilla RNN units, trained for 14 epochs
4	<i>BiRNN-ELMo</i>	one layer of 50 vanilla RNN units, trained for 13 epochs
5	<i>BiLSTM-GloVe</i>	one layer of 32 bidirectional LSTM units, trained for 31 epochs
6	<i>BiLSTM-ELMo</i>	one layer of 64 bidirectional LSTM units, trained for 14 epochs
7	<i>BiGRU-GloVe</i>	one layer of 64 bidirectional GRUs, trained for 14 epochs
8	<i>BiGRU-ELMo</i>	one layer of 64 bidirectional GRUs, trained for 8 epochs

Table 3.2: Hyperparameter values used in the 8 sub-models

3.6.2 Ensemble Classifiers

Scikit-learn (Pedregosa et al., 2011) was used to develop the SVM classifier used in the ensemble model. Three different versions of ensemble classifiers were developed:

- (1) ***Ens-Feat*** is the version of the ensemble model that only utilizes the neural features. The SVM classifier in this version uses a sigmoid kernel. The γ and C parameters in the SVM were set to *auto* (i.e., $1/\langle \text{number-of-features} \rangle$) and 4, respectively.
- (2) ***Ens-Prob*** uses only the predicted class probabilities from the softmax activation function at the end of the neural sub-models. It utilizes a polynomial kernel with the degree of 1. The γ and C parameters in the SVM were set to *scale* (i.e., $1/[\langle \text{number-of-features} \rangle \times \langle \text{variance-of-features} \rangle]$) and 1, respectively.
- (3) ***Ens-All*** utilizes both neural features and predicted class probabilities in its SVM classifier, that uses a sigmoid kernel, and has its values of γ and C set to *auto* and 2, respectively.

3.6.3 Submitted Runs

Based on the results on the validation set, 5 runs were submitted to the shared task server. For the 1st and 2nd runs, *CNN-GloVe* and *CNN-ELMo* were used, respectively, as stand-alone models³, and *Ens-Feat*, *Ens-Prob*, and *Ens-All* comprised the 3rd, 4th and 5th runs.

³These two sub-models achieved the most promising results among all the sub-models, during the training phase.

3.7 Shared Task Results

Table 3.3 shows the official results of our submissions, as well as selected runs from other teams (as reported in Losada et al., 2019) that achieved the best result in terms of one of the official evaluation metrics, or achieved competitive results. For the results of our team (*CLaC*), we indicate in Table 3.3 the specific name of the models used in the five submitted runs.

team	model	run	#writings	P	R	F1	ERDE ₅	ERDE ₅₀	l-w F1
CLaC	CNN-GloVe	0	109	.45	.74	.56	6.72%	3.93%	.54
CLaC	CNN-ELMo	1	109	.61	.82	.70	5.73%	3.13%	.69
CLaC	Ens-Feat	2	109	.60	.81	.69	6.02%	3.13%	.68
CLaC	Ens-Prob	3	109	.63	.76	.69	6.27%	3.55%	.68
CLaC	Ens-All	4	109	.64	.79	.71	6.25%	3.43%	.69
lirmm		0	2024	.74	.63	0.68	9.13%	5.14%	.63
lirmm		1	2024	.77	.60	.68	9.10%	5.51%	.62
Fazl		2	2001	.09	1	.16	17.11%	11.22%	.14
UNSL		0	2000	.42	.78	.55	5.54%	3.92%	.55
UNSL		4	2000	.31	.92	.47	6.14%	2.97%	.46
INAOE-CIMAT		3	2000	.67	.68	.68	9.17%	4.57%	.63

Table 3.3: Official results on the first sub-task of the eRisk 2019 shared task. *#writings*: maximum number of writings (Reddit posts) that were processed for a user, *P*: Precision, *R*: Recall, *l-w F1*: Latency-Weighted F1 score.

As shown in Table 3.3, the model *Ens-All* achieved the highest F1 (0.7073) and latency-weighted F1 (0.6908) scores of all participants’ runs. This was in line with our intuition that using an ensemble model that makes use of both neural features and predicted class probabilities from the 8 sub-models has a higher capability of detecting the correct class after observing a small number of writings. The results also show that the *CNN-ELMo* model can achieve F1 and latency-weighted F1 scores that are competitive to *Ens-All*, and outperforms *Ens-Feat* and *Ens-Prob* based on these two metrics. The *CNN-ELMo* model also resulted in the best recall, *ERDE*₅ and *ERDE*₅₀, showing the potential of this model to be used independently for the task of early risk detection of anorexia.

Table 3.3, also shows that all our models, except *CNN-GloVe* (run 0) achieved significantly superior performances in terms of F1 score and latency-weighted F1 (teams *lirmm* and *INAOE-CIMAT* achieved the next best F1 and latency-weighted F1 scores). Run 1 of team *lirmm* achieved the highest precision. The best recall was achieved by run 2 of the team *Fazl*. Runs 0 and 4 of the team *UNSL* achieved the highest *ERDE*₅ and *ERDE*₅₀, respectively, where we could achieve

competitive results using the *CNN-ELMo* model.

Finally, by comparing the results achieved by *CNN-GloVe* and *CNN-ELMo* in Table 3.3, we can observe the advantages of using ELMo to extract context-sensitive embeddings in achieving a better performance.

Although our systems achieved the best or competitive results according to different evaluation metrics, our implementation of the inference system could have been more efficient. In our system, at every iteration, all the available posts were processed again by the model. As the release of the test data was in a one by one fashion, our approach resulted in a time complexity which was quadratic to the number of posts. Considering that the majority of the processing time in our system was happening in running the ELMo embedder, this approach was not efficient enough to handle 2000 iterations in a timely manner.

Knowing that the post-level-representation computation component of our model (i.e., everything before the user-level attention component) was decoupled across posts, this aspect could have been leveraged in order to have the post representations calculated only once and cached. Using this approach, at each iteration, the representations of the newly available posts would have been computed, and would have been passed to the user-level attention along with the retrieved post representations from the previous iterations. This would have reduced the time complexity to linear, making it more efficient to be able to handle all the 2000 iterations.

3.8 Explainability Analysis

As a further analysis, we investigated the usefulness of the user-level attention component as a means to explain our model’s behavior. Due to achieving comparable results to the best performing model, i.e., the *Ens-All*, we focused our experiments on the *CNN-ELMo* model. Furthermore, in order to have an exclusive focus on explainability, early prediction was not included as a factor in these experiments. As a result, testing was done by utilizing the full test set.

3.8.1 Experiments

For our new set of experiments, first, we passed the posts for each user in the test set to the previously trained *CNN-ELMo* model. Utilizing the weights assigned by the user-level attention component of this model to every user’s different posts, we ranked the posts of each user from the highest-weighted to the lowest-weighted. Then we monitored the performance of our model in the following two scenarios:

- 1) When feeding only the n top weighted posts to the model: the n was gradually increased from 1 (i.e., only the top-weighted post from a user) to 1000.
- 2) When doing scenario 1, but by having the user-level attention mechanism being replaced with an average pooling layer.

Using the 2 scenarios above, we were able to monitor the performance of the model when only top-weighted posts from each user were provided. We could also evaluate the effect of the user-level attention through ablation.

3.8.2 Results and Discussion

The results from the experiments are shown visually in Figure 3.2, and the original numerical data is provided in Table 3.4.

Looking at the solid lines in Figure 3.2, it can be observed that increasing the number of available posts per user will generally lead to improvement in terms of accuracy, precision, and F1 of the model with the user-level attention (scenario 1), while leading to a drop in recall. The solid lines also show that, as the number of available posts increases, the performance measures gradually converge to their final values, which is when all the posts by users are available. We hypothesize that the gradual improvement in precision and drop in recall is due to the behavior of the user-level attention that was used to rank a user’s writings by assigning weights to posts based on the amount of signal indicating a user being anorexic. As a result, when a low number of highly-weighted posts is passed to the model, it has a higher tendency to predict the author as anorexic leading to higher recall. We also believe that the convergence in the performance measures is because the lowest-weighted posts

Max # of Posts/User	with Attention (Scenario 1)				with Avg Pool (Scenario 2)			
	A	P	R	F1	A	P	R	F1
1	65.15	19.60	93.15	32.38	65.15	19.60	93.15	32.38
2	69.57	21.86	93.15	35.42	72.39	23.43	91.78	37.33
3	74.23	25.09	94.52	39.65	77.55	27.64	93.15	42.63
4	77.18	27.49	94.52	42.59	80.61	30.59	91.78	45.89
5	79.88	29.96	93.15	45.33	83.19	32.61	82.19	46.69
6	80.74	30.37	89.04	45.30	85.28	35.40	78.08	48.72
7	81.84	31.53	87.67	46.38	86.63	38.00	78.08	51.12
8	82.33	32.16	87.67	47.06	87.24	39.16	76.71	51.85
9	83.56	33.86	87.67	48.85	87.85	40.44	75.34	52.63
10	84.29	34.97	87.67	50.00	88.83	43.08	76.71	55.17
20	88.59	43.06	84.93	57.14	91.90	54.02	64.38	58.75
30	90.18	47.29	83.56	60.40	93.74	67.74	57.53	62.22
40	92.27	54.46	83.56	65.95	93.50	69.23	49.31	57.60
50	93.13	58.09	83.56	68.54	93.37	71.11	43.84	54.24
60	93.74	61.00	83.56	70.52	93.50	75.00	41.10	53.10
70	94.23	63.54	83.56	72.19	93.01	72.22	35.62	47.71
80	94.48	64.89	83.56	73.05	92.76	71.87	31.51	43.81
90	94.85	67.03	83.56	74.39	92.76	75.00	28.77	41.58
100	95.21	69.32	83.56	75.78	92.64	74.07	27.40	40.00
200	96.07	76.62	80.82	78.67	92.76	88.89	21.92	35.16
300	96.19	77.63	80.82	79.19	92.27	85.71	16.44	27.59
400	96.20	77.63	80.82	79.19	92.39	92.31	16.44	27.91
500	96.32	80.28	78.08	79.17	92.27	91.67	15.07	25.88
600	96.56	82.61	78.08	80.28	92.02	90.00	12.33	21.69
700	96.56	82.61	78.08	80.28	91.90	88.89	10.96	19.51
800	96.44	81.43	78.08	79.72	91.90	88.89	10.96	19.51
900	96.81	85.07	78.08	81.43	91.90	88.88	10.96	19.51
1000	96.69	83.82	78.08	80.85	91.90	88.88	10.96	19.51
2000 (all)	96.93	86.36	78.08	82.01	91.90	88.89	10.96	19.51

Table 3.4: Performance of the model (in percentage) in terms of the maximum number of highly-weighted posts from each user that have been used during inference. The columns titled *with Average Pooling* belong to the model in which the user-level attention mechanism is ablated. The last row belongs to the case when all the posts from each user are provided for the system.

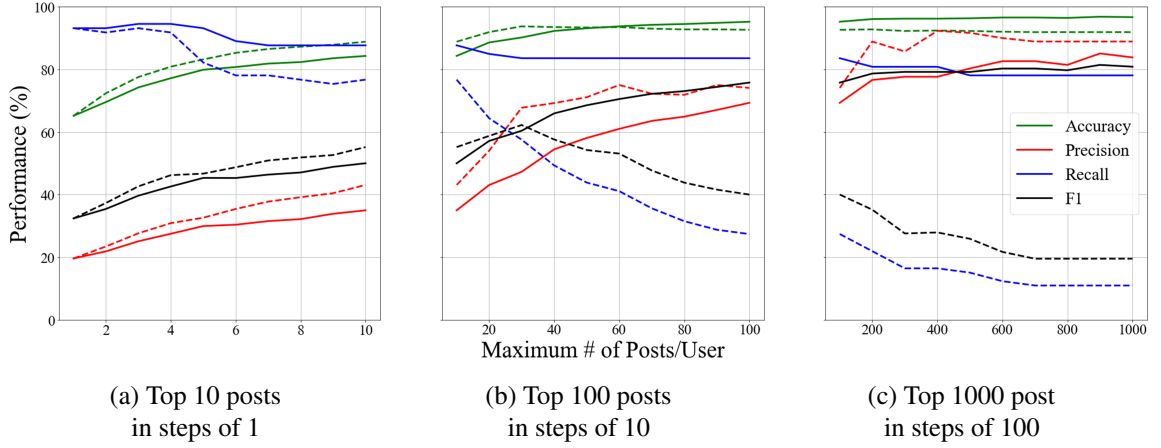


Figure 3.2: Performance of the model in terms of the maximum number of highly-weighted posts from each user that have been provided during inference. The solid lines correspond to the model with user-level attention (scenario 1), while the dotted lines correspond to the model with user-level average pooling (scenario 2).

play a very small role in the model’s prediction.

As the dotted lines in Figure 3.2a show, increasing the maximum number of available posts from 1 to 10 will lead to gradual improvements in accuracy, precision, and F1, while causing a drop in the recall. It is hypothesized that, in general, the 10 highest-weighted posts include the necessary information for the model to make a prediction. It can also be observed that, the model with average pooling even achieves better improvements in terms of F1 and accuracy than the model with the user-level attention. We believe this is because the model with the user-level attention tends to get biased towards specific posts.

Looking at Table 3.4 and Figure 3.2b, it can be observed that, starting at 30 posts, the F1 score achieved by the model with the average pooling starts to deteriorate. The same behavior happens to the accuracy of this model at 50 posts. Comparing the results with the ones achieved by the model with the user-level attention, it can be observed that the model with the user-level attention is more capable in handling a higher number of posts.

Finally, the rapid drop in the recall of the model with average pooling when increasing the number of available posts can be observed. This shows that, with a higher number of posts available for a user, this model loses its capability to recognize the patterns that are useful in the detection of anorexia. This observation is also aligned with the hypothesis that the weights assigned to posts by

the user-level attention generally correlate with the level of signals for anorexia that the posts emit.

In order to further validate the explainability of the of user-level attention, we studied the values of the highest weights assigned by the attention mechanism. We focused on two categories of users: predicted by the model as positive (i.e., true positive and false positive), and predicted by the model as negative (i.e., true negative and false negative). For each category, we averaged the n -th highest weights assigned to the posts by the user-level attention. The results for this analysis is provided in Table 3.5.

Rank	W_p	W_n	W_p/W_n
1 st	0.388	0.056	6.96
2 nd	0.124	0.034	3.66
3 rd	0.067	0.026	2.62
4 th	0.047	0.021	2.18
5 th	0.035	0.019	1.86
6 th	0.029	0.017	1.70
7 th	0.024	0.015	1.59
8 th	0.019	0.014	1.31
9 th	0.016	0.013	1.23
10 th	0.015	0.012	1.20

Table 3.5: Average weights assigned by the user-level attention mechanism to the n^{th} highest weighted posts for users detected by the system as anorexic (W_p) or non-anorexic (W_n)

As indicated in Table 3.5, in case users are predicted by the model as anorexic, the weights assigned to the highest weighted posts are, on average, 6.96 times greater in comparison to the users detected by the model as non-anorexic. This magnitude drops in lower-ranked posts and gradually converges to 1. This could convey that, the model is generally more likely to predict the anorexic label for a user, if the user-level attention mechanism assigns a high weight to a post by this user. It can also indicate that, in cases where the model makes a positive prediction, it has heavily relied on the evidence that it finds from a limited number of posts which it deems as important. This observation is also in support of the hypothesis that the weights assigned to posts by the user-level

attention correlate with the amount of anorexia signals that the model finds in posts.

3.8.3 Further Analysis

We performed a further analysis to evaluate contribution of the n-grams to the importance weights assigned to each post by the attention mechanism. Considering that the model utilizes unigram and bigram CNN encoders, we know that the model only uses these as initial information sources; hence, we focus our study on these n-grams only. Our approach is described as follows:

We start by considering only one convolution component with k kernels with a specific window size n . If it is applied over a post p containing N n-grams, and the output is fed to a CReLU layer, the output of the CReLU y would have the size $N \times 2k$. We consider y_i as the i -th column in y ($i \in [1 \dots 2k]$), hence the length of y_i is equal to N . Equation 9 shows the formula to trace where the maximum element in y_i (t_i^p) is computed:

$$t_i^p = \arg \max(y_i) \quad (9)$$

Then, once the post representation is calculated by applying a max pooling to the output of the CReLU, we compute an *element-level contribution value* ECV_i^p by multiplying the element by its corresponding weight within the attention layer. Having ω_i as the attention layer's weight corresponding to the i -th element in the post representation, ECV_i^p is computed using Equation 10.

$$ECV_i^p = \max(y_i) \times \omega_i \quad (10)$$

We then compute a *post-level contribution value* for each unique n-gram W within the post p . For each W , the post-level contribution value PCV_W^p is computed by summing the ECV_i^p s for the values of i where the (t_i^p) th n-gram is equal to W :

$$PCV_W^p = \sum_{i, w_{(t_i^p)}=W} ECV_i^p \quad (11)$$

The final *contribution value* for n-gram W is computed by averaging the PCV_W^p values across posts that include this n-gram. The process is shown in Equation 12:

$$CV_W = \frac{1}{|P_W|} \sum_{p \in P_W} PCV_W^p \quad (12)$$

We ran the aforementioned method for both the unigram and bigram CNN encoders, and then ranked all the unigrams and bigrams based on their contribution values. Our results show that 99 out of the first 100 ranked items are unigrams containing the names of mental or physical illnesses. The first rank belonged to the word “anorexia”, and it was followed by “leukaemia”, “dystonia”, “leukemia”, “preeclampsia”, “poliosis”, “anemia”, “dyslexia”, “osteosarcoma”, and “insomnia”. It is also interesting to find that the misspelled word “annorexia” was ranked 23rd based on its contribution value. This is mainly because of ELMo’s good handling of morphological variants, which does not let the misspelled words be considered out of vocabulary, but rather generates a meaningful embedding for them.

It was also observed that 85 out of the top 100 bigrams had “weight” as their second token. The top bigram was “gain weight”, and it was followed by “appropriate weight”, “gained weight”, “major weight”, “versus weight”, “muscle weight”, “gaining weight”, “good weight”, “unwanted weight”, “healthy weight”, and so on. Overall, the unigrams were given higher ranks than the bigrams, with the highest-ranked bigram, “gain weight” showing up in the 81st rank. We also observed that the 15 top bigrams among the top 100, which did not include the word “weight” seemed generally irrelevant to the task (e.g., “fight fight”, “video comparison”, or “4-d chess”) and may have merely been considered important by the model, due to their frequent occurrence within posts by anorexic users in the training data, and do not seem informative on their own.

Overall, the following observations can be made:

- The model considers the mentioning of the names of mental or physical illnesses as very important clues.
- The model relies more on unigrams than bigrams. This is supported by the overall higher contribution values of unigrams in comparison to the bigrams, the very shallow scope of the top bigrams, and the existence of possibly irrelevant bigrams among the top ones.

Observing the model’s narrow scope of focus on specific unigrams and bigrams and keeping in

mind the same model was the most-performing deep learning model in the task of early detection of anorexia, it is observed that the task of detection of anorexia is an n-gram triggered task in most cases and not much contextual information is required in order to make a correct prediction. However, this also implies that the model is biased towards specific unigrams and bigrams, and may potentially make a wrong prediction merely based on observing those unigrams and bigrams in someone’s social media posts (e.g., posts such as “I do not have *anorexia*.” or “I don’t care about *gaining weight*.” leading the model towards predicting a user as anorexic). We believe that the use of recurrent models in the ensemble approach has helped to some extent in overcoming this limitation with the *CNN-ELMo* model, and has resulted in a slightly better performance in the *Ens-All* model. But we believe that further work needs to be done in order to develop a model that can better leverage contextual information, while still paying enough attention to the triggering n-grams.

3.9 Conclusion

This chapter described our participation to the first sub-task of eRisk 2019, and presented the ensemble approach we developed to detect anorexia in social media. The ensemble model utilized neural features alongside predicted class probabilities which are output by 8 different neural sub-models. Using this model and under the team name *CLaC*, our submission ranked first in terms of F1 and latency-weighted F1 scores in the first sub-task of eRisk 2019.

This chapter also extended our paper at eRisk 2019 (Mohammadi et al., 2019b), and presented the experiments that were done in order to evaluate the explainability of the model. The experimental results showed that the the user-level attention mechanism assigned weights to the posts based on the amount of anorexia signals present in posts. This capability could be leveraged in a real-life use case to provide a limited set of most signalling posts in order for the experts in clinical psychology to validate the model’s predictions.

We also discovered that, despite the model’s high performance, it has limitations in terms of capturing useful information at the intra-post level, as it mainly focuses on the existence of certain unigrams and bigrams. A future direction for this research would be to develop a model that better captures contextual information in posts, while also not missing out on the important n-grams that

are relevant in the detection of anorexia.

Chapter 4

Quantitative Measurement of Attention Faithfulness

Chapter 3 showed that attention can highlight post-level evidence in early risk detection, but it also raised the question of whether these weights truly reflect the model’s decision process. In this chapter, we take a step further and propose a quantitative way to measure attention faithfulness. Our goal is to move beyond qualitative observations and provide a reproducible metric that can be used to compare models and training setups. More specifically, we address the second research question introduced in Section 1.3:

Can attention weights provide faithful explanations in text classification tasks, and can their faithfulness be quantified reliably?

The work presented in this chapter was published in the proceedings of the 2022 Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackBoxNLP 2022) (Amini & Kosseim, 2022).

Although attention weights have been commonly used as a means to provide explanations for deep learning models, this approach has been widely criticized due to its lack of *faithfulness*. In this chapter, we propose a new metric called ATTEFA to quantitatively represent the degree of faithfulness of attention weights. Using this metric, we further validate the effect of the frequency of informative input elements and the use of contextual vs. non-contextual encoders on the faithfulness of the

attention mechanism. Finally, we apply the approach on several real-life binary classification datasets to measure the faithfulness of attention weights in real-life settings.

Our findings show that attention faithfulness has an inverse correlation with the degree of encoder contextualization and the frequency of informative tokens. We also observed that single-headed attention is usually not faithful in real-life binary classification tasks.

4.1 Introduction

Attention mechanism (Bahdanau et al., 2015) has become an indispensable part of many state-of-the-art Natural Language Processing (NLP) models, and its application is becoming more and more prevalent in non-NLP use cases. In simple words, and from a functionality perspective, attention can be described as a module which generates outputs from the representations of input elements by performing the following two steps:

- (1) Automatically compute weights corresponding to each input element;
- (2) Use the computed weights to run a weighted average over the input representations.

Due to the attention’s explicit mechanism to assign weights to input elements, attention weights have been frequently used as explanations for model predictions. A common approach has been to provide attention heat maps to show which input elements the attention component has attended to (e.g., Y. Wang, Huang, Zhu, and Zhao, 2016; Lee, Shin, and Kim, 2017; Lin et al., 2017; Ghaeini, Fern, and Tadepalli, 2018). However, the use of attention weights as explanations has been widely challenged, with regards to the observation that they are not *faithful*, meaning that different attention weights can result in similar model predictions (Jain & Wallace, 2019; Serrano & Smith, 2019; Wiegrefe & Pinter, 2019). Therefore, the explanations provided by the attention weights are neither unique, nor closely related.

In this work, we extend the work of Wiegrefe and Pinter (2019) to a single-run adversarial setup that can be used to compute a quantitative metric for the faithfulness of attention weights, called ATTEFA (*Attention Faithfulness*). We consider the adversarial training setup single-run in the sense that it can provide the ATTEFA metric by running the adversarial training only once.

To perform a sanity check on ATTEFA, we run experiments in a controlled setting using synthetic datasets and two types of encoders, i.e., a non-contextual Multi-Layer Perceptron (MLP) and a contextual Long Short-Term Memory (LSTM), to help us validate if the values of this metric reflect what we expect it to be. We later compute this metric on real-life binary text classification datasets to validate how faithful the attention weights are in those settings, and conclude that single-headed attention is usually not faithful in real-life binary text classification settings.

4.2 Related Work

Since the rise of deep learning models, researchers have focused on devising techniques that could provide an explanation for the functioning of these so-called “black-box” models. Among different classes of explainability techniques, the following are the most notable.

Gradient-based methods attribute model decisions to input features using gradient signals ([Aubakirova & Bansal, 2016](#); [Karlekar, Niu, & Bansal, 2018](#); [Selvaraju et al., 2017](#); [Sundararajan, Taly, & Yan, 2017](#)). Perturbation-based methods try to provide an explanation for the model behavior by evaluating its reactions to perturbations in the input features ([Ribeiro et al., 2016](#); [Zintgraf, Cohen, Adel, & Welling, 2017](#)). Attention-based methods act as an intuitive way of interpreting the model’s decision. They use the probability distribution or weights provided by an attention mechanism as a feature importance measure to find the features that the model is attending to ([Luong, Pham, & Manning, 2015](#); [Mullenbach, Wiegrefe, Duke, Sun, & Eisenstein, 2018](#); [Xie, Ma, Dai, & Hovy, 2017](#)).

Despite the popularity of the attention-based explainability approaches, the reliability of these methods has been called into question, with a particular focus on the faithfulness of the explanations provided by the attention mechanism. [Jain and Wallace \(2019\)](#) performed different experiments to evaluate the meaningfulness of explanations provided by attention weights. Their results show that attention weights are not correlated with gradient-based feature importance scores. Furthermore, they show that it is often possible to have different attention probability distributions that result in a similar output, arguing that a specific distribution cannot be treated as the definitive cause behind a model decision.

[Serrano and Smith \(2019\)](#) investigated the ability of attention weights to act as importance

measures through a different lens. They stated that it is not sufficient for the weights to make sense to humans. The weights should also provide a faithful explanation for the model output in order to be considered reliable. Through performing multi-weight tests, they show that although there is a certain level of correlation between attention weights and the importance of features in the final prediction of the model, these weights in many cases cannot successfully identify the features that heavily impact a model’s decision.

Wiegrefe and Pinter (2019) proposed additional tests for evaluating the ability of the attention mechanism to provide explainability. They challenged the findings reported by Jain and Wallace (2019) as they treated the attention as a stand-alone component within a network that is independent from the rest of the components. Using an end-to-end adversarial setup, they trained models to produce similar outputs while forcing different attention distributions in binary classification tasks, and showed that attention-based explanations were not as unfaithful as Jain and Wallace (2019) had suggested.

Wiegrefe and Pinter (2019) introduced a scalar hyperparameter, λ , to balance the divergence of the attention distributions against the divergence of the output predictions. They iterated over different values of λ and computed the adversarial model’s attention Jensen-Shannon Divergence (JSD) and prediction Total Variation Distance (TVD), presenting them in a two-dimensional plot to qualitatively assess whether the resulting curves were convex (indicating unfaithful attention) or not. We argue that with minor adjustments to how the divergences of attention weights and output predictions are computed, it is possible not only to eliminate the need for λ , but also to develop a quantitative metric for attention faithfulness.

In our work, we extended the adversarial setup of Wiegrefe and Pinter (2019) through re-scaling of attention weights and output prediction divergence values, so that:

- (1) The need for the tuning of the hyperparameter λ is removed.
- (2) The adversarial training step requires only a single run per model (as opposed to Wiegrefe and Pinter’s setup which required multiple runs with different values of λ).
- (3) The faithfulness of the attention can be measured in a quantitative rather than a qualitative way.

This approach results in a metric, which we call ATTEFA, that can provide a quantitative insight on how faithful the explanations by the attention component are, given a specific model and a specific dataset. To the best of our knowledge, this is the first work that provides such a quantitative measure to evaluate the faithfulness of attention.

4.3 Method

In this section, we describe the methodology for computing ATTEFA for binary text classification tasks. The approach is to first train a base attention-based model to perform the task, and then train an adversarial model to mimic the predictions of the base model but using a different attention distribution. The ATTEFA can then be computed using the prediction and attention weight divergences between the base and the adversarial models.

4.3.1 Base Model Training

First, we train a base model on the binary text classification data. In our experiments, the base model is comprised of an embedding layer, followed by an encoder (LSTM or MLP), which is in turn followed by an attention component, and finally a classification head. To train the base model, cross-entropy loss is used, and training is done for 8 epochs. The final base model is the trained checkpoint at the end of the epoch where the ROC-AUC score on the test dataset is minimum. This setup was selected similarly to [Wiegrefe and Pinter \(2019\)](#).

4.3.2 Adversarial Model Training

With the base model at hand, we train an adversarial model with the same architecture as the base model, but with the following two characteristics:

- (1) Having predictions as similar as possible to the base model, and
- (2) Having attention weight distributions as different as possible from the base model

Following [Jain and Wallace \(2019\)](#) and [Wiegrefe and Pinter \(2019\)](#), in order to measure the difference between the adversarial and the base models' predictions, namely $\hat{y}_a$ and $\hat{y}_b$, we use TVD,

which is computed using Equation 13:

$$\text{TVD}(\hat{y}_a^j, \hat{y}_b^j) = \frac{1}{2} \sum_{j=1}^{|\mathcal{Y}|} |\hat{y}_a^j - \hat{y}_b^j| \quad (13)$$

where $|\mathcal{Y}|$ represents the number of output heads (which is equal to 1 in our binary classification setting).

Following Jain and Wallace (2019) and Wiegrefe and Pinter (2019), to compute the difference between attention distributions α_a and α_b , JSD is used, which is computed using Equation 14:

$$\text{JSD}(\alpha_a, \alpha_b) = \frac{1}{2} \text{KLD}(\alpha_a || \bar{\alpha}) + \frac{1}{2} \text{KLD}(\alpha_b || \bar{\alpha}) \quad (14)$$

where $\bar{\alpha} = \frac{\alpha_a + \alpha_b}{2}$ and the Kullback-Leibler Divergence (KLD) is computed using Equation 15:

$$\text{KLD}(\alpha_a || \alpha_b) = \sum_{k=1}^{|\alpha|} \alpha_a^k \times \left(\log(\alpha_a^k + \epsilon) - \log(\alpha_b^k + \epsilon) \right) \quad (15)$$

where $|\alpha|$ corresponds to the size of the attention weight vector. The inclusion of ϵ in the KLD equation is to prevent the logs from becoming infinite in cases where the values of α become equal to zero due to mathematical underflow. In our experiments, we set the value of ϵ equal to 1e-10.

Having the TVD of the predictions and the JSD of the attention weight distributions, we design the loss function so that it tries to minimize TVD and maximize JSD. The final loss formula is given in Equation 16:

$$\mathcal{L}(\mathcal{M}_a, \mathcal{M}_b)^{(i)} = \text{sTVD}(\hat{y}_a^{(i)}, \hat{y}_b^{(i)}) - \text{sJSD}(\alpha_a^{(i)}, \alpha_b^{(i)}) \quad (16)$$

In Equation 16, we use sTVD and sJSD to denote the scaled values of TVD and JSD, respectively, both bounded between 0 and 1. We apply the scaling in order to make sure that the value ranges for the TVD and JSD components of the loss are equal, and therefore the final value of the loss is affected equally by the two components. Knowing that the value of TVD is always between 0 and 0.5, sTVD is computed using Equation 17:

$$\text{sTVD}(\hat{y}_a, \hat{y}_b) = \frac{\text{TVD}(\hat{y}_a, \hat{y}_b)}{0.5} \quad (17)$$

To compute sJSD, Equation 18 is used:

$$\text{sJSD}(\alpha_a, \alpha_b) = \frac{\text{JSD}(\alpha_a, \alpha_b)}{\text{JSD}_{\max}} \quad (18)$$

where $\text{JSD}_{\max}$ is the calculated upper-bound for JSD when Equations 14 and 15 are used. $\text{JSD}_{\max}$ is approximately equal to 0.6931, and is reached when α_1 and α_2 in Equation 14 are two one-hot vectors with the element 1 located in different indices.

The value of $\mathcal{L}(\mathcal{M}_a, \mathcal{M}_b)^{(i)}$ is computed per sample i . In order to compute the backpropagated loss value for each batch, we compute the average over the per-sample losses in the batch:

$$\mathcal{L}(\mathcal{M}_a, \mathcal{M}_b) = \frac{1}{N} \sum_{i=1}^N \mathcal{L}(\mathcal{M}_a, \mathcal{M}_b)^{(i)} \quad (19)$$

where N is the batch size.

The training process is continued until the loss value on the test data does not improve for 10 consecutive epochs, or a maximum number of 80 epochs is reached. To calculate the total loss on the test data, instead of computing the per-sample losses and averaging them over the dataset, for simplicity and to leverage the metric implementations by [Wiegrefe and Pinter \(2019\)](#), we first average over the per-sample TVD and JSD on the test dataset, and then compute the total loss using these averages. As the final adversarial model, we pick the checkpoint from the training epoch with the lowest value of loss on the test data.

The key difference between our adversarial training setup and the one from [Wiegrefe and Pinter \(2019\)](#) is in the way the adversarial loss is computed. In [Wiegrefe and Pinter \(2019\)](#), KLD is used instead of JSD to compute the distribution divergence between the base model’s attentions and the adversarial one. Since the value of KLD is un-bounded, it is mandatory to use an additional hyperparameter λ to avoid the final value of the loss getting dragged fully towards the attention divergence. Knowing that JSD has a specific lower and upper bound, including that in the adversarial loss formula allows us to do away with the additional hyperparameter λ , and to be able to do the adversarial training in one shot, which in turn provides us with an easy and systematic way to compute a metric value for the attention faithfulness.

4.3.3 Computing ATTEFA

Having the TVD of the predictions and the JSD of the attention distributions on the test data between the base model $\mathcal{M}_b$ and adversarial version of the model $\mathcal{M}_a$, we compute the faithfulness score ATTEFA of the attention module $\mathcal{A}_{\mathcal{M}}$ using Equation 20:

$$\text{ATTEFA}(\mathcal{A}_{\mathcal{M}}) = \min \left(\frac{\text{sTVD}(\hat{y}_a, \hat{y}_b)}{\text{sJSD}(\alpha_a, \alpha_b)}, 1 \right) \quad (20)$$

The formula is motivated by the assumption that, the degree of attention faithfulness has a direct relation with the value of the TVD of the predictions, and an inverse relation with the value of the JSD of the attention weights. In other words, if the attention is faithful, meaning that the attention can find a limited set of informative sources, the adversarial setup will either converge to a point where both the TVD of predictions and the JSD of the attention weights are low, or both of them are high. We believe that the second scenario is more probable, as the adversarial model has a much higher degree of freedom in order to converge to a different attention distribution from the base model than to achieve a similar output prediction. It will later be shown in Section 4.6 that, with the current adversarial setup, the adversarial model usually achieves a JSD close to its maximum value.

With this assumption, we believe that in most cases, the final value for $\text{sTVD}(\hat{y}_a, \hat{y}_b)$ should be lower than $\text{sJSD}(\alpha_a, \alpha_b)$, but we still do not rule out the opposite scenario, which is why we force the value of ATTEFA to be bounded between 0 and 1 through the use of the min function in Equation 20.

4.4 Datasets

In order to validate the use of ATTEFA, we experimented with a collection of binary text classification datasets.

The first set is comprised of multiple synthetic sentiment analysis datasets, which are used to sanity check the computation approach for ATTEFA. These synthetic datasets were developed specifically for this work to assess the impact of the frequency of informative tokens. In Section 4.4.1, we explain the motivation behind using these datasets and the methodology to develop them.

The second set includes the datasets used in the work of [Jain and Wallace \(2019\)](#) and [Wiegrefe](#)

and Pinter (2019). The motivations for using these datasets were to measure the faithfulness of single-headed attention in real-life use cases, along with the ability to compare our findings with Wiegrefe and Pinter (2019). Section 4.4.2 provides a brief description of these datasets.

4.4.1 Synthetic Datasets

We hypothesize that, in principle, the faithfulness of the attention has a direct relation with the rareness of the informative elements in the input. In the task of text classification, we assume that attention weights should be assigned to each input token proportional to their informativeness to the task. If there are very few informative tokens, our hypothesis is that the attention should focus on those and not the other tokens, and finding alternative attention weight distributions that would lead to a similar outcome would be difficult. On the other hand, in cases where many input tokens are informative and helpful to the task, the attention can easily shift its focus from one set of tokens to another, therefore the faithfulness will be low.

In order to verify this hypothesis, we designed a set of synthetic sentiment analysis datasets that include different proportions of informative texts. To that end, we synthetically created samples in a way that a specific portion of the tokens are words with the same sentiment polarity as the sentiment label of the sample¹, while filling the rest of the token slots with the uninformative token "something", whose polarity is considered neutral. This results in a simple-to-classify sentiment dataset that allows us to investigate the effect of the frequency of informative input elements on the faithfulness of attention, without the need to take into account the effectiveness of attention for the task at hand.

Our *Mock* datasets are comprised of 8000 training and 1000 testing samples. The distribution of the positive/negative labels is 50/50 in the datasets, and each sample has a random length between 50 and 100 tokens. These synthetic datasets are comprised of the following:

- **Mock-1:** Only one token in each sample is informative.
- **Mock-2:** 2 tokens in each sample are informative.

¹We picked words with positive and negative sentiment from the following gazetteers, respectively: <https://ptrckpry.com/course/ssd/data/positive-words.txt>, <https://ptrckpry.com/course/ssd/data/negative-words.txt>

- **Mock-5:** 5 tokens in each sample are informative.
- **Mock-10:** 10 tokens in each sample are informative.
- **Mock-1q:** 25% of the tokens in each sample are informative.
- **Mock-2q:** 50% of the tokens in each sample are informative.
- **Mock-3q:** 75% of the tokens in each sample are informative.
- **Mock-4q:** All (i.e., 100% of) the tokens in each sample are informative.

Appendix [A.1](#) shows examples from these datasets.

4.4.2 Real-life Datasets

Following the work of [Jain and Wallace \(2019\)](#) and [Wiegreffe and Pinter \(2019\)](#), we benchmarked our approach using the following real-life datasets:

- **Diabetes:** This is the *Diabetes* subset of the MIMIC IDC9 dataset ([Johnson et al., 2016](#)). The dataset samples are the discharge summaries of electronic health records, and the task is to determine if the summaries have been labeled as diabetic or not.
- **Anemia:** This is the *Chronic vs Acute Anemia* subset of the MIMIC IDC9 dataset ([Johnson et al., 2016](#)). Similar to *Diabetes*, the dataset samples are discharge summaries, but here they are annotated as *acute* (0) vs. *chronic* (1), depending on the type of anemia that the patient was diagnosed with.
- **SST:** The original *Stanford Sentiment Treebank* dataset ([Socher et al., 2013](#)) contains sentences tagged with 5 different sentiment values, ranging from most negative (class 1) to most positive (class 5). To convert this into a binary classification task, [Jain and Wallace \(2019\)](#) filtered out neutral samples (class 3), and split the remaining into 2 buckets, one for negative (classes 1 and 2) and the other for positive (classes 4 and 5).
- **IMDB:** The *IMDB Large Movie Reviews Corpus* ([Maas et al., 2011](#)) contains movie reviews annotated according to their sentiment polarity (positive vs. negative).

- **AgNews:** The *AG News Corpus* (Del Corso, Gulli, & Romani, 2005; Gulli, 2005) contains news articles from more than 2000 sources. For the binary classification task, Jain and Wallace (2019) only considered the two categories of *world* (class 0) and *business* (class 1).
- **20News:** The *20 Newsgroups* dataset (Lang, 1995) contains newsgroup correspondences for 20 categories. For this dataset, Jain and Wallace (2019) only considered the classes *baseball* (class 0) and *hockey* (class 1) in their binary classification setting.

4.4.3 Dataset Statistics

Table 4.1 shows the average number of tokens across samples, along with the distribution of the positive/negative samples for each dataset. Since all the synthetic datasets include the same number of samples, class distributions, and average number of tokens across samples, we have included the statistics for them under *Mock*-*.

Dataset	Train		Test	
	Size (neg/pos)	Avg Len (Tokens)	Size (neg/pos)	Avg Len (Tokens)
Mock-*	4000/4000	75	500/500	75
Diabetes	6650/1416	1985	1389/340	2385
Anemia	1742/2912	2368	512/857	2396
SST	3310/3610	17	912/909	17
IMDB	8673/8539	180	2189/2174	176
AgNews	25508/25492	36	1900/1900	36
20News	612/624	159	192/195	206

Table 4.1: Summary statistics of the datasets.

4.5 Experimental Setup

As discussed in Section 4.3, in our work, two families of models are experimented with: *LSTM*-based and *MLP*-based.

The model architectures were selected similarly to [Wiegrefe and Pinter \(2019\)](#). The *LSTM* models are comprised of the following components:

- (1) A 300d word embedding layer.
- (2) A bidirectional LSTM layer ([Hochreiter & Schmidhuber, 1997](#)) with 128 units.
- (3) The single-head attention module.
- (4) A fully-connected layer.

The *MLP* models include embedding, attention, and fully-connected modules similar to the *LSTM* models, but utilize a feed-forward projection layer with 128 nodes followed by a *tanh* activation, instead of the BiLSTM layer.

The attention has a two layer fully-connected network that first projects the input to half its size in its first layer, applies a *tanh* activation, and then maps it to a single logit in the second layer. A softmax function is then used to convert the logit to a probability distribution, which is used to compute a weighted average over the inputs and form the output of the attention.

Similarly to [Jain and Wallace \(2019\)](#) and [Wiegrefe and Pinter \(2019\)](#), for the Diabetes and Anemia datasets, 300d Word2Vec embeddings ([Mikolov et al., 2013](#)) are pre-trained on the combined text from the two datasets. The training is done using CBOW with a window size of 10. For the rest of the datasets, 300d publicly-pretrained FastText embeddings ([Bojanowski, Grave, Joulin, & Mikolov, 2017](#)) are used.

Adam ([Kingma & Ba, 2015](#)) is used as the optimizer during training, and the learning rate and weight decay rates are set to 1e-3 and 1e-5, respectively. Weight decay is applied to every component in the network except the attention module.

4.6 Results and Discussion

Table 4.2 shows the F1 scores achieved by the base models (see Section 4.3.1). In order to verify the correctness of our experiments, we have also included in the table the F1 scores reported by [Wiegrefe and Pinter \(2019\)](#).

Dataset	LSTM		MLP	
	Reported	Reproduced	Reported	Reproduced
Mock-1	-	0.974	-	0.975
Mock-2	-	0.988	-	0.989
Mock-5	-	0.999	-	1.000
Mock-10	-	1.000	-	0.999
Mock-1q	-	1.000	-	1.000
Mock-2q	-	1.000	-	1.000
Mock-3q	-	1.000	-	1.000
Mock-4q	-	1.000	-	1.000
Diabetes	0.775	0.733	0.699	0.665
Anemia	0.938	0.935	0.920	0.915
SST	0.831	0.830	0.817	0.816
IMDB	0.902	0.908	0.888	0.882
AgNews	0.964	0.959	-	0.956
20News	0.942	0.935	-	0.878

Table 4.2: F1 scores of the base model achieved on the test datasets. The F1 scores reported by [Wiegrefe and Pinter \(2019\)](#) have been included under the *Reported* columns. The MLP setup is equivalent to the *Trained MLP* setup from [Wiegrefe and Pinter \(2019\)](#).

Table 4.3 contains the results achieved by the adversarial setup. It includes the F1 scores of the adversarial models, the TVD of their predictions from the base models, the JSD of their attention distributions from the base models, the number of epochs that resulted in the best loss on test, and their attention faithfulness score ATTEFA. The numbers are reported in terms of average and standard deviation runs with different random seeds. Individual results for each seed is available in Tables A.9 and A.10 in Appendix A.2.

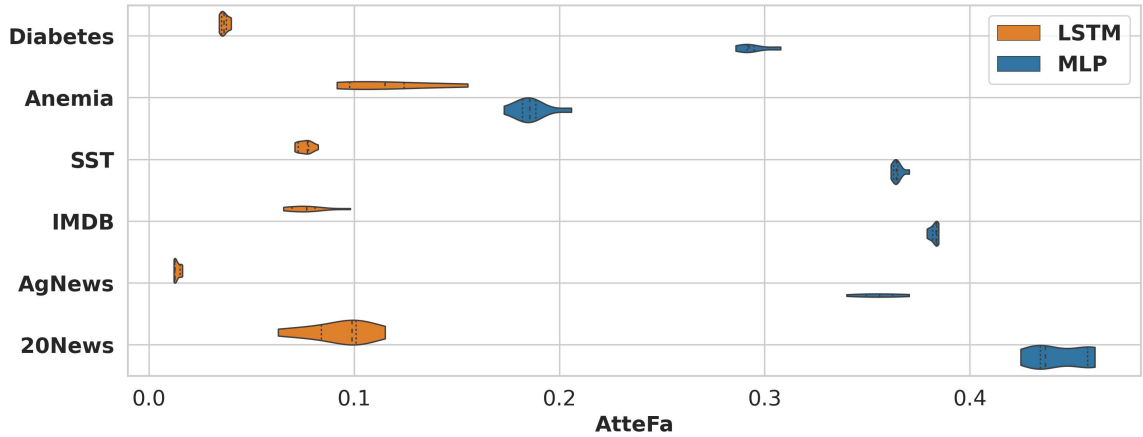


Figure 4.1: Distribution of ATTEFA across different models and real-life datasets.

Dataset	LSTM					MLP				
	Epoch	F1	TVD	JSD	ATTEFA	Epoch	F1	TVD	JSD	ATTEFA
Mock-1	12±5	0.947±0.020	0.015±0.009	0.693±0.000	0.0304±0.0176	20±24	0.221±0.312	0.231±0.012	0.393±0.000	0.8153±0.0425
Mock-2	9±7	0.977±0.010	0.016±0.005	0.693±0.000	0.0329±0.0096	1±0	0.221±0.312	0.246±0.003	0.670±0.000	0.5086±0.0064
Mock-5	7±5	1.000±0.001	0.002±0.000	0.693±0.000	0.0037±0.0008	9±11	0.147±0.275	0.247±0.001	0.686±0.000	0.4987±0.0027
Mock-10	14±8	1.000±0.000	0.001±0.000	0.693±0.000	0.0012±0.0000	23±31	0.147±0.275	0.249±0.001	0.686±0.000	0.5028±0.0027
Mock-1q	23±12	1.000±0.000	0.000±0.000	0.693±0.000	0.0006±0.0000	37±31	0.465±0.480	0.135±0.120	0.689±0.001	0.2721±0.2407
Mock-2q	35±32	1.000±0.001	0.000±0.000	0.678±0.004	0.0008±0.0010	42±35	0.751±0.324	0.099±0.109	0.691±0.000	0.1983±0.2188
Mock-3q	21±21	1.000±0.000	0.000±0.000	0.681±0.008	0.0003±0.0001	13±13	0.999±0.001	0.001±0.001	0.691±0.000	0.0013±0.0011
Mock-4q	8±4	1.000±0.000	0.000±0.000	0.680±0.004	0.0002±0.0003	3±0	1.000±0.000	0.000±0.000	0.690±0.000	0.0002±0.0000
Diabetes	22±5	0.729±0.003	0.018±0.001	0.693±0.000	0.0367±0.0020	42±27	0.134±0.076	0.147±0.004	0.691±0.000	0.2945±0.0072
Anemia	20±6	0.901±0.018	0.058±0.011	0.693±0.000	0.1164±0.0211	23±10	0.832±0.007	0.093±0.004	0.692±0.000	0.1861±0.0083
SST	21±6	0.823±0.002	0.034±0.002	0.626±0.006	0.0760±0.0034	23±15	0.605±0.028	0.173±0.001	0.656±0.002	0.3645±0.0024
IMDB	49±12	0.889±0.006	0.038±0.004	0.691±0.001	0.0769±0.0090	21±14	0.158±0.056	0.190±0.001	0.689±0.000	0.3826±0.0019
AgNews	49±18	0.958±0.001	0.007±0.001	0.683±0.002	0.0136±0.0015	24±12	0.610±0.032	0.172±0.005	0.671±0.001	0.3558±0.0097
20News	18±5	0.865±0.013	0.046±0.007	0.689±0.001	0.0931±0.0149	24±18	0.340±0.149	0.208±0.004	0.650±0.008	0.4444±0.0133

Table 4.3: Average and standard deviation of the results from our adversarial setup. The results for every row are reported from 9 different runs with different random seed initializations. The column *epoch* indicates the training epoch at which the selected checkpoint was taken for evaluation.

4.6.1 Effect of Contextualization

Comparing the ATTEFA columns for the *LSTM* and *MLP* models in Table 4.3, we can observe that the attentions incorporated in models with LSTM as their encoder are significantly less faithful than their counterparts in the models with MLP as their encoder. This observation was not surprising, as a lower degree of contextualization in token representations should inherently result in higher faithfulness in the attention that is applied on top of those representations.

To better understand this, imagine the task of detecting whether a text is about sports or fruits. If you want to classify the example `football is life`, the only informative word in the example is `football`, as it clearly indicates a sport. In an ideal scenario, a faithful attention should have a distribution highly centered on this word. Using an MLP encoder, the input tokens will retain their information, therefore the representation of token `football` retains its informativeness. This is, however, not necessarily the case if a contextual encoder such as an LSTM is used to compute the token representations, as the LSTM can manipulate the representations in a way that another word, such as `is`, has the informative representation.

Going back to our adversarial setting, when an LSTM is used, the encoder has the capacity to manipulate the token representations so that a different set of tokens bear the useful information to achieve the task. In this setting, the attention can simply focus on the new set and obtain similar information. On the other hand, a non-contextual MLP encoder does not have the capacity that an LSTM holds, and will retain the informativeness of the representation for each token. Therefore, it becomes more challenging for the attention to find a new set of tokens to attend to. That is why the prediction TVD in the MLP models is significantly lower than the LSTM ones, resulting in the MLP models having a noticeably higher ATTEFA.

Simply put, our results show that attention components applied on top of contextual encoders are generally less faithful than the ones on top of non-contextual encoders.

4.6.2 Effect of the Frequency of Informative Sources

Looking at the rows corresponding to the results on the *Mock*-* datasets and the MLP model in Table 4.3, we can observe the general trend towards the reduction of ATTEFA as the number of

informative tokens increase. For the case of the MLP model, a relatively high ATTEFA of ≈ 0.82 is achieved on the Mock-1 dataset, which only includes one informative token in each sample text, and the value drops to close to 0 for the case of Mock-3q and Mock-4q datasets. This shows that the faithfulness of the attention mechanism has an inverse correlation with the number of informative sources in the input.

The trend is still observable in the case of the LSTM models, but with a magnitude that is considerably lower than what we have for the MLP models, as the ATTEFA on the Mock-1 dataset is only ≈ 0.03 . As discussed in Section 4.6.1, the contextualized LSTM encoder has the flexibility to re-distribute the task-relative information across different input tokens. Regardless of that, we can still observe the general trend towards the drop of ATTEFA as we move from Mock-1 to Mock-4, which shows that, even with the case of contextualization, the frequency of informative elements in the source input can still affect the faithfulness of the attention mechanism.

We can observe anomalies in the trend described above. For example, we can observe increases in the ATTEFA in Mock-1 to Mock-2 and Mock-1q to Mock-2q for the case of the LSTM model, and from Mock-5 to Mock-10 in the case of the MLP model. This can be partially justified by the behavior of the base model in terms of how successful it is in detecting informative tokens. An example of this can be found in Table 4.2, where the MLP model has achieved a slightly lower F1 score on Mock-10 (1.000) in comparison to Mock-5 (0.999), meaning that the attention used in the MLP model was more successful in identifying informative tokens in the Mock-5 dataset than in Mock-10.

We can also observe a $\approx 18\%$ gap between the ATTEFA of the MLP model trained on the Mock-1 dataset (≈ 0.82) and the maximum value of ATTEFA (i.e., 1). We argue that this is also related (at least partially) to how the base model performs. We can see in Table 4.2 that the base model trained on the Mock-1 dataset does not have an F1 score of 1 on the test dataset (0.975). This could partially be due to the failure of attention to detect the informative tokens and highly focus on them.

Overall, we conclude that there is generally an inverse relation between the frequency of informative sources in the input data and the faithfulness of the attention module trained on it. But there is still some noise in the ATTEFA metric which is attributed to how well the base model performs. Although we do not believe that this rules out ATTEFA as a suitable metric to compute the

faithfulness of attention, we believe there is room for exploring alternative metrics that, for example, also incorporate the performance of the base models in their computation.

4.6.3 ATTEFA on Real-life Datasets

Table 4.3 shows that for the case of the MLP models, the values of ATTEFA on all the real-life datasets are significantly lower than the ones on Mock-1 to Mock-10. As discussed in Section 4.6.2, this could show that there is quite a large number of informative tokens in the samples belonging to these datasets, which allows the attention to shift its focus among them. This shows that, the attention mechanism in MLP models trained on all these datasets is not very faithful.

For the case of the LSTM model, however, we can observe that the ATTEFA on these real-life datasets is comparable and sometimes higher than their counterparts on the Mock-* datasets. However, focusing only on the real-life datasets, the ATTEFA of the LSTM models are still lower than the MLP ones. This can also be visually observed in Figure 4.1, which includes the violin plots of the distribution of ATTEFA across the different datasets and models. We suspect that, in real-life datasets, we have a significantly lower number of completely uninformative tokens as we had in the Mock-* datasets. Although the LSTM encoder still retains its flexibility to redistribute information across different tokens, the lower number of completely uninformative tokens reduces the degree of the information redistribution capacity. This is something that we have not explored in our experiments with the synthetic datasets, and therefore, leaves room for more studies on this aspect.

One may argue that the number of input tokens on its own can affect the distribution of attention weights and can in turn affect the value of the attention JSD of the adversarial models, hence the final value of ATTEFA. While we do not rule this out, we believe that it is not merely the input lengths that would affect the attention JSD, but rather the frequency of informative input tokens that could increase as the input lengths become higher. We also believe that the way information is distributed among their representations used by the attention component also plays a big role here.

In Figure 4.1, we can see that for the case of the MLP models, the values of ATTEFA on datasets with lengthier samples, namely Diabetes and Anemia, are generally lower than the ones on the other datasets. This is, however, not the case for the LSTM models, as we can observe a relatively high

ATTEFA on the Anemia dataset with respect to the rest of the datasets. Even for the case of the MLP model, we can see that the ATTEFA on the 20News dataset is higher than SST and AgNews that have lower average input lengths (see Table 4.1).

We therefore conclude that the distribution of task-related information across the input token representations used by the attention component plays a key role in the faithfulness of the attention.

4.6.4 Comparison of Our Adversarial Setup with Wiegrefe and Pinter’s

In Figure 4.2, we have plotted the prediction TVD and attention JSD of our adversarial LSTM models against the results reported in Wiegrefe and Pinter (2019). The dotted lines in the plots reproduce the ones in Figure 5 from Wiegrefe and Pinter (2019).

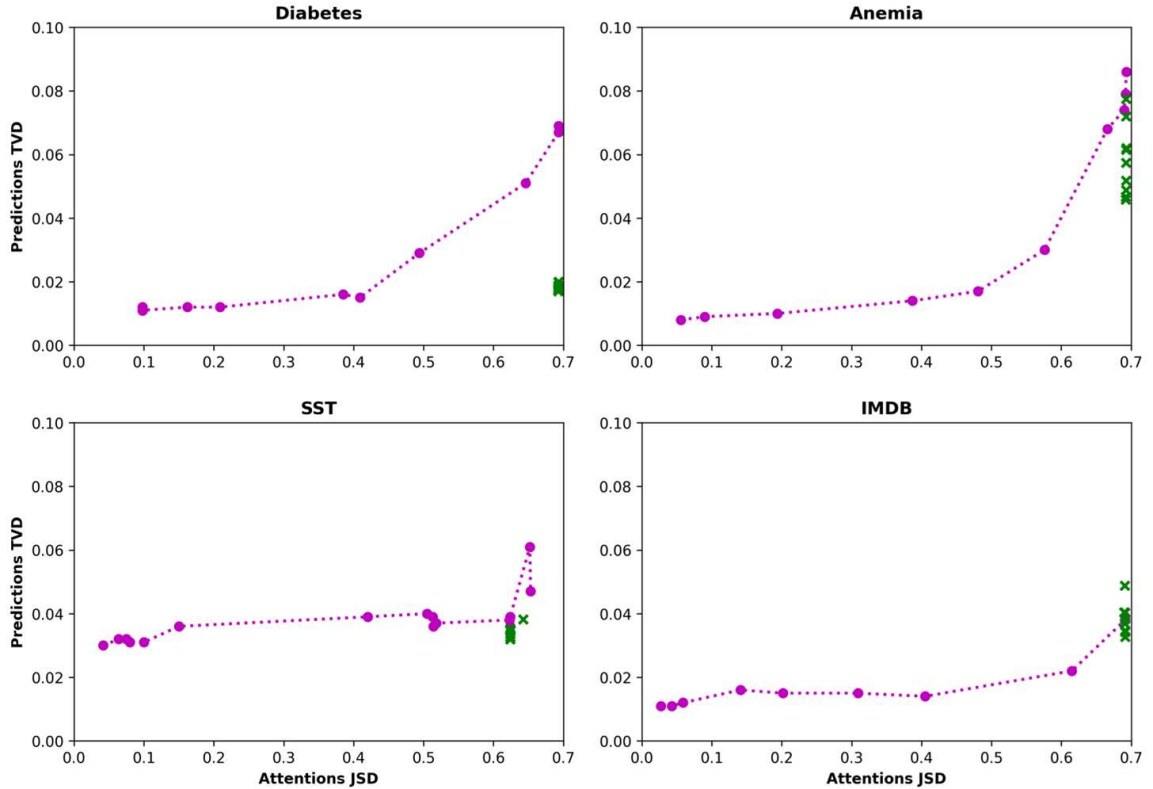


Figure 4.2: Visual comparison of averaged per-instance test set JSD and TVD from base model for each model variant between our adversarial setup and the one from Wiegrefe and Pinter (2019). The $\bullet$ show results from Wiegrefe and Pinter (2019), and the $\times$ show results from our setup.

As Figure 4.2 shows, with our adversarial setup, we have achieved comparable prediction TVDs to Wiegrefe and Pinter’s on the Anemia, SST and IMDB datasets. However, on the Diabetes dataset,

our prediction TVDs are significantly lower than Wiegreffe and Pinter’s. Given that our adversarial setups are similar, we believe that this is mainly due to our inability to properly reproduce their base LSTM model on the Diabetes dataset. We can observe this from the 0.042 drop in the F1 score of our model from what was reported in Wiegreffe and Pinter (2019) (see Table 4.2).

Figure 4.2 also shows that the adversarial results that we have achieved are towards the higher-end of the attention JSDs reported by Wiegreffe and Pinter (2019). This is very close to the calculated upper-bound for JSD, which is 0.6931. Wiegreffe and Pinter used the hyperparameter λ in order to reduce the effect of the attention JSD in the value of their loss. With the removal of this hyperparameter in our setup (which is the equivalent of setting it to 1), the adversarial training leads the model to primarily maximize the attention JSD, as it is an easier objective than to minimize the prediction TVD. Therefore, this leads to an almost maxed-out attention JSD, and it is mainly the prediction TVD that determines the value of ATTEFA. However, we argue that the JSD is not always fully maxed-out (see the plot for the SST dataset in Figure 4.2), and therefore, we cannot simply disregard it in the computation of ATTEFA.

4.7 Conclusion

In this chapter, we proposed an adversarial training approach for binary text classification tasks, which can provide us with the metric ATTEFA that quantitatively measures the degree of faithfulness in the attention weights. We, then, measured the effect of contextualization, as well as the effect of the frequency of informative tokens on the attention faithfulness. Finally, we computed and evaluated ATTEFA for models trained on several real-life binary text classification datasets.

There are certain limitations with the current work, in terms of both the methodology used to compute ATTEFA, and the different factors affecting the attention faithfulness. In this section, we present the most prominent future work.

The current methodology to compute ATTEFA is scoped solely on binary text classification. In order to make ATTEFA more widely applicable, the methodology needs to be extended to other NLP tasks, such as multi-class classification, text retrieval, question answering, machine translation, etc.

In the current work, we have studied the effect of the frequency of informative tokens on the

faithfulness of attention through running experiments on the Mock-* datasets, which are synthetic datasets for sentiment classification. The current selection of sentiment words and their positioning within the input texts were done in a random fashion. A more thorough experiment would explore the effect of the distribution of informative tokens across the input texts (centered towards the start/end/middle vs. scattered evenly), along with a more careful selection of the words to be used as the informative tokens (e.g., differentiating between words with strong vs. weak sentiments).

In terms of investigating the effect of encoder contextualization on the faithfulness of attention, we have explored using token-level MLP as a non-contextual encoder and LSTM as a contextual one. This can be extended to exploring other encoder architectures, such as CNNs (LeCun et al., 1999), GRUs (Cho et al., 2014), and transformers (Vaswani et al., 2017).

Another aspect in the current study which has room for exploration is the evaluation of the effect of softmax temperature on the faithfulness of attention. In this work, we used the standard softmax function with a temperature of 1. We believe that higher faithfulness may be achieved by using lower temperatures in the case of datasets with infrequent informative tokens, and higher temperature in the case of datasets with frequent informative tokens within their input.

Last, but not least, the experiments in this work are only focused on single-head attention. We believe that the current approach does not transfer well to multi-headed attentions, as we may still consider a multi-headed attention faithful if the only way for the adversarial model to arrive at the same predictions as the base model is to change the order of the attention heads and not the attention weights. Due to the frequent use of multi-headed attentions in state-of-the-art NLP models, the extension of ATTEFA to multi-headed attentions is an important research avenue.

Chapter 5

User-Aware Attention for Explainable Detection of Mental Health Risks in Social Media

In this chapter, we address the third research question introduced in Section 1.3:

Can we develop an approach that simultaneously achieves competitive early risk detection and produces explanations with demonstrated faithfulness?

This chapter builds on the modeling work of Chapter 3 and the evaluation perspective of Chapter 4. From Chapter 3, we keep the idea of using attention as post-level evidence for early detection on user timelines. From Chapter 4, we bring in a more systematic way of testing whether those explanations are faithful. The contents of this chapter were published in the proceedings of the 2025 International Conference on Recent Advances in Natural Language Processing (RANLP 2025) (Amini & Kosseim, 2025).

In this chapter, we propose a user-aware attention-based framework for early detection of mental health risks from social media posts (modernized and strengthened compared to the model in Chapter 3). Our model combines DISORBERT (Aragon et al., 2023), a mental health–adapted transformer encoder, with a user-level attention mechanism that produces transparent post-level explanations. To assess whether these explanations are faithful, i.e., aligned with the model’s

true decision process, we apply adversarial training and quantify attention faithfulness using the ATTEFA metric (see Chapter 4). Experiments on four eRisk tasks (depression, anorexia, self-harm, and pathological gambling) show that our model achieves competitive latency-weighted F1 scores while relying on a sparse subset of posts per user. We also evaluate attention robustness and conduct ablations, confirming the model’s reliance on high-weighted posts. Our work extends prior explainability studies by integrating faithfulness assessment in a real-world high-stakes application. We argue that systems combining predictive accuracy with faithful and transparent explanations offer a promising path toward safe and trustworthy AI for mental health support.

5.1 Introduction

Mental health disorders are major public health concerns, and early identification of conditions such as depression, anorexia, self-harm, and pathological gambling is critical. As social media becomes a widespread platform for self-expression, online posts have emerged as valuable data sources for detecting mental health risks. Prior shared tasks, notably the *eRisk* challenges, have facilitated research on early risk prediction from Reddit posts, encouraging models that can issue timely and accurate alerts for at-risk individuals (Losada et al., 2017, 2018, 2019, 2020; Parapar et al., 2021, 2022, 2023, 2024).

However, deploying such systems in practice requires not just predictive accuracy but also *trustworthy explanations*. Black-box decisions can be problematic in sensitive domains like mental health, where users and clinicians need to understand the reasoning behind predictions. Attention mechanisms have been widely used in neural Natural Language Processing (NLP) models as a means of offering transparency, producing weights over input segments that are often interpreted as explanations.

Yet, recent work has called this interpretability into question. Jain and Wallace (2019) showed that attention weights are often not faithful, i.e., they may not align with the model’s actual decision logic. Wiegrefe and Pinter (2019) argued that attention can still be useful if properly validated. The distinction between *plausibility* (i.e., what seems intuitive to humans) and *faithfulness* (what the model actually uses) has since become central in explainability research (Jacovi & Goldberg, 2020;

[Moradi et al., 2021](#)). In high-stakes settings like mental health, faithfulness is essential.

In this work, we propose a user-aware attention-based framework for early detection of mental health risks on social media. Our model combines DISORBERT ([Aragon et al., 2023](#)), a BERT model adapted to mental health and social media domains, with a user-level attention mechanism that identifies the most informative posts as explanations. To evaluate the faithfulness of these explanations, we further employ adversarial training to evaluate attention faithfulness using ATTEFA ([Amini & Kosseim, 2022](#)), a metric that quantifies how accurately the attention weights reflect the true importance of input features in the model’s actual decision-making process.

We validate our framework on four eRisk tasks (depression, anorexia, self-harm, and pathological gambling), achieving competitive latency-weighted F1 scores. Our analysis confirms that the model typically relies on a sparse subset of posts to make predictions, highlighting both the rarity of strong mental health signals and the value of faithful explanations in this domain.

Our main contributions are as follows:

- We propose a novel user-aware attention-based architecture for early detection of mental health risks from social media, integrating DISORBERT with user-level attention for post-level explainability.
- We quantitatively evaluate the faithfulness of user-level attention weights through adversarial training and the use of the ATTEFA metric.
- We conduct extensive experiments on four eRisk shared tasks (depression, anorexia, self-harm, and pathological gambling), achieving competitive latency-weighted F1 scores and demonstrating that only a small subset of posts are typically required for accurate classification.

5.2 Related Work

5.2.1 Mental Health Detection from Social Media

Early work demonstrated the feasibility of using linguistic and behavioral patterns from social media to detect mental health conditions ([Coppersmith et al., 2014, 2015](#); [De Choudhury et al., 2013](#)). The CLEF eRisk shared tasks ([Losada et al., 2017, 2018, 2019, 2020](#); [Parapar et al., 2021](#),

2022, 2023, 2024) formalized early detection tasks across several disorders, emphasizing timely decision-making over Reddit data.

More recent approaches leverage transformer-encoder-based models, fine-tuned on Reddit or X datasets. [Murarka et al. \(2021\)](#) used RoBERTa ([Zhuang, Wayne, Ya, & Jun, 2021](#)) to detect multiple mental disorders with strong performance. [Aragon et al. \(2023\)](#) introduced DISORBERT, a variant of BERT ([Devlin et al., 2019](#)) adapted first to social media language and then to mental health content, showing improved accuracy across the eRisk datasets.

User-level modeling has also gained traction, often using hierarchical architectures to aggregate post representations. Hierarchical Attention Networks (HANs) ([Z. Yang et al., 2016](#)) allow models to weight each post, highlighting the most indicative content. [Zogan, Razzak, Wang, Jameel, and Xu \(2022\)](#) proposed a multi-aspect HAN that augmented attention-based signals with user metadata, improving both interpretability and performance for depression detection.

While these models provide some degree of transparency through their built-in attention mechanism, few studies have systematically assessed whether explanations reflect the model’s decision process. This gap motivates a deeper examination of attention *faithfulness* in the detection of mental health risks using social media.

5.2.2 Explainability and Attention Faithfulness

Attention mechanisms are often viewed as providing interpretable insights into neural predictions. However, [Jain and Wallace \(2019\)](#) showed that attention distributions can be manipulated without affecting model outputs, questioning their explanatory power. [Wiegrefe and Pinter \(2019\)](#) argued that attention can serve as explanation if validated using diagnostic tests, including adversarial attention training.

Other studies proposed ways to evaluate *faithfulness* of the attention mechanism. [Serrano and Smith \(2019\)](#) and [H. Chen, Zheng, and Ji \(2020\)](#) used input ablation to measure the impact of attended tokens. [Chrysostomou and Aletras \(2021\)](#) introduced Task-Specific Scaling to bias attention weights toward task-relevant tokens. More recently, [Amini and Kosseim \(2022\)](#) proposed ATTEFA, a metric that quantifies how well attention weights reflect a model’s prediction under adversarial perturbations.

Our work builds on these efforts by enforcing faithfulness through adversarial training and evaluating explanation integrity via ATTEFA. We extend attention-based interpretability to the user level in a real-world task, showing that faithful explanations can enhance trust in systems for mental health support.

5.3 Methodology

In this section, we describe our user-aware attention-based framework for early detection of mental health risks from social media. The model operates at two levels: a post-level encoder generates contextualized embeddings for each user post, and a user-level attention mechanism aggregates these embeddings to produce a prediction, while simultaneously providing post-level interpretability. We employ DISORBERT as the base encoder, leveraging its domain adaptation to social media and mental health language. To ensure that attention-based explanations are faithful to the model’s decision process, we apply adversarial training and quantify explanation faithfulness using the ATTEFA metric. We also outline our training setup, adversarial objective, datasets, and evaluation protocol.

In relation to the methodology presented in Chapter 3, our new approach builds on novel resources and knowledge that were not available when the model in Chapter 3 was developed. Specifically:

- (1) We replace CNN/RNN over GloVe/ELMo with DISORBERT and fine-tune it end-to-end together with the user-level attention and classification head. This shift was motivated by the maturation of a stable transformer fine-tuning frameworks, i.e., the HuggingFace *Transformers* library (Wolf et al., 2020), along with the access to stronger GPUs¹, which were not available when Chapter 3 was developed.
- (2) We adopt more efficient training choices: a $\tanh$ non-linearity in the user-level attention, AdamW (Loshchilov & Hutter, 2019) instead of Adam (Kingma & Ba, 2015), and a more systematic use of regularization, e.g., dropout and weight decay, the latter available directly through the default configuration of AdamW in PyTorch (Paszke et al., 2019).

¹A100 GPUs were provided as part of Concordia’s CSSE GPU cluster.

Full architectural and training details are provided in Sections 5.3.1 and 5.3.2.

5.3.1 Model

Figure 5.1 shows the overall architecture of the model, together with the inputs and outputs for each layer. The model classifies each user as *at risk* or *without risk* based on its social media posts. In the following subsections, we will review each layer separately.

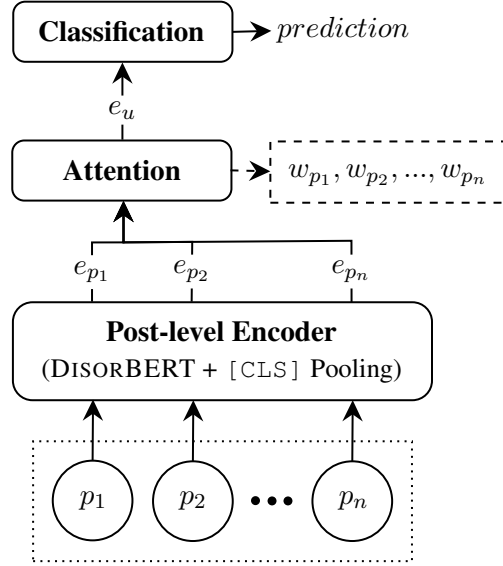


Figure 5.1: Model Architecture. $p_1, p_2, \dots, p_n$ represent the posts by user u , e_{p_i} and w_{p_i} denote the embedding and the attention weight for the post p_i , and e_u is the final embedding of the user u that is output by the attention.

Post-level Encoder

The post-level encoder layer is responsible for computing the 1D embedding (e_{p_i}) for each individual post by the user (p_i).

As the encoding model, we used DISORBERT (Aragon et al., 2023), which is a BERT-based model that is domain-adapted to both the mental health domain and social media language. We applied first pooling to extract the representation of the [CLS] token as the embedding for each post (e_{p_i}).

User-level Attention

Given a user u with posts $p_1, p_2, \dots, p_n$, we first compute the logit l_{p_i} for each post p_i using Equation 21:

$$l_{p_i} = \text{F2}(\tanh(\text{F1}(e_{p_i}))) \quad (21)$$

where e_{p_i} is the embedding for the post p_i , computed using the post-level encoder (see Section 5.3.1). F1 is a fully connected layer, with $|e_{p_i}|$ input and $|e_{p_i}|/2$ output nodes. F2 is also a fully connected layer, with $|e_{p_i}|/2$ input nodes and 1 output node².

Having the logits for all posts, the attention weight w_{p_i} for each post p_i is computed through softmax, as shown in Equation 22:

$$w_{p_i} = \frac{\exp(l_{p_i})}{\sum_{j=1}^n \exp(l_{p_j})} \quad (22)$$

Lastly, the embedding e_u for the user u is computed through a weighted average over $e_{p_1}, e_{p_2}, \dots, e_{p_n}$, with $w_{p_1}, w_{p_2}, \dots, w_{p_n}$ as weights:

$$e_u = \sum_{i=1}^n w_{p_i} e_{p_i} \quad (23)$$

The attention layer returns both e_u (fed to the classification head), along with the vector $[w_{p_1}, w_{p_2}, \dots, w_{p_n}]$, which we use to determine the informativeness of each post.

Classification Head

The classification head is composed of a dropout (with $p = 0.1$), followed by a feed-forward layer with one output node, followed by a sigmoid activation. Having e_u as the input, this component is used to predict the likelihood of the user u being *at risk* for the corresponding mental health problem.

²In our experiments, $|e_{p_i}|$ was equal to 768 (i.e. the hidden size of DISORBERT).

5.3.2 Training Process

In this section, we describe the training procedures for the base and adversarial models. The base model is trained to detect mental health risks from user posts. In contrast, the adversarial model is specifically optimized to replicate the base model’s predictions while assigning different attention weights to the user’s posts. This approach enables a rigorous evaluation of whether the attention mechanism faithfully reflects the model’s true decision process.

Base Model

Training is done using PyTorch (Paszke et al., 2019) and HuggingFace *Transformers* (Wolf et al., 2020). We initialize the *Post-level Encoder* from the pre-trained checkpoint of DISORBERT. The rest of the parameters in the model are randomly initialized.

Training is done using binary cross-entropy as the loss function. While the loss is back-propagated on a per-sample basis, we use gradient accumulation to leverage an effective batch size of 4. In order not to overflow the GPU memory, we enable gradient checkpointing (T. Chen, Xu, Zhang, & Guestrin, 2016), and feed the posts in minibatch sizes of 32. The text for each post is created through joining the title and the body of the corresponding post with a newline character (`\n`)³.

AdamW (Loshchilov & Hutter, 2019) is used as the optimizer. We set the learning rate to $1e - 5$, and use the default values in Pytorch for the remaining optimizer hyperparameters.

Training is continued until the area under the precision-recall curve (PR-AUC) over the test dataset does not improve in 10 consecutive epochs, or a maximum of 50 epochs is reached. After the completion of training, the trained model checkpoint is picked at the epoch where the maximum PR-AUC over the test dataset was reached. We then set the prediction threshold so that it maximizes the F1 score on the test data.

Adversarial Model

We follow a similar approach to Amini and Kosseim (2022) (see Chapter 4) to train the adversarial model. Using the trained base model, we compute the prediction and the user-level attention weights

³To accelerate training, we limited each post to its first 200 tokens.

for each user in the dataset.

For each sample, the prediction loss is initially calculated as the absolute difference between the outputs of the base and adversarial models:

$$\mathcal{L}_y^i = |y_a^i - y_b^i| \quad (24)$$

where y_a^i and y_b^i denote the outputs of the adversarial and the base models, respectively, on sample i .

We then compute the Jensen-Shannon Divergence (JSD) between the user-level attention weights between the adversarial and the base models. With α_a^i and α_b^i representing the attention weights of the adversarial and the base models, respectively, for sample i , and $\bar{\alpha}_i = \frac{\alpha_a^i + \alpha_b^i}{2}$, the JSD between α_a^i and α_b^i is computed using Equation 25:

$$\text{JSD}(\alpha_a^i, \alpha_b^i) = \frac{1}{2} \text{KLD}(\alpha_a^i || \bar{\alpha}_i) + \frac{1}{2} \text{KLD}(\alpha_b^i || \bar{\alpha}_i) \quad (25)$$

To compute the Kullback-Leibler Divergence (KLD), we use $\epsilon = 1e-10$ to avoid a mathematical error when computing the $\log(0)$. With this ϵ , the maximum possible value for JSD is ~ 0.6931 , which we will later refer to as JSD_{max} .

We then compute the second portion of the loss, using the following equation:

$$\mathcal{L}_\alpha^i = \frac{\text{JSD}(\alpha_a^i, \alpha_b^i)}{\text{JSD}_{max}} \quad (26)$$

The division by JSD_{max} ensures the same theoretical range (i.e., between 0 and 1) for both $\mathcal{L}_y^i$ and $\mathcal{L}_\alpha^i$. With $\mathcal{L}_y^i$ and $\mathcal{L}_\alpha^i$ at hand, the adversarial loss for sample i is computed using Equation 27:

$$\mathcal{L}^i = \mathcal{L}_y^i - \mathcal{L}_\alpha^i \quad (27)$$

During training, the final loss to be back-propagated for each batch is computed by averaging the individual values of adversarial loss ($\mathcal{L}^i$) for each sample within the batch.

We use the average adversarial loss over the test data as the early stopping criterion. Aside from the computation of loss and the early stopping criterion, the remaining setup for the adversarial

training phase is similar to the training of the base model.

Amini and Kosseim (2022) observed that maximizing the JSD between the user-level attention weights of the adversarial and the base model is a simpler task compared to minimizing the distance between the outputs of the two models. We argue that this would increase the chances of convergence to a local minima, where the $\mathcal{L}_y^i$ component of the loss is not fully optimized. To avoid such a problem, we have performed the adversarial training from both random initialization and the trained base checkpoints, with the earlier case using a similar setup as Amini and Kosseim (2022). In the latter scenario, $\mathcal{L}_y^i$ starts from a value close to zero⁴, which would theoretically guide the adversarial model to retain close predictions to the trained base model.

5.3.3 Datasets and Tasks

We evaluate our setup using the following 4 datasets from the eRisk shared task series (Losada et al., 2017, 2018, 2019, 2020; Parapar et al., 2021, 2022, 2023). The datasets contain a collection of reddit posts and comments⁵, grouped by users and sorted in chronological order. Each user is labelled as *at risk* or *without risk* for the corresponding mental health issue, using the methodology proposed in Losada and Crestani (2016). Table 5.1 shows the sources for each dataset, and Table 5.2 shows the distribution of the number of positive (i.e., *at risk*) and negative (i.e., *without risk*) samples (users) in the training and test datasets for each task.

⁴Due to the use of regularization techniques during training, the value of $\mathcal{L}_y^i$ is not exactly zero.

⁵For simplicity, the term “post” is used to encompass both posts and comments.

Dataset	Subset	Source(s)
Anorexia	Train	eRisk 2018 - T2 (Losada et al., 2018)
	Test	eRisk 2019 - T1 (Losada et al., 2019)
Depression	Train	eRisk 2017 - T1 (Losada et al., 2017) eRisk 2018 - T1 (Losada et al., 2018)
	Test	eRisk 2022 - T2 (Parapar et al., 2022)
Gambling	Train	eRisk 2021 - T1 (Parapar et al., 2021) eRisk 2022 - T1 (Parapar et al., 2022)
	Test	eRisk 2023 - T2 (Parapar et al., 2023)
Self-harm	Train	eRisk 2019 - T2 (Losada et al., 2019) eRisk 2020 - T1 (Losada et al., 2020)
	Test	eRisk 2021 - T2 (Parapar et al., 2021)

Table 5.1: Sources of the datasets.

Dataset	Train		Test	
	pos	neg	pos	neg
Anorexia	61	411	73	742
Depression	214	1,493	98	1,302
Gambling	245	4,182	103	2,071
Self-harm	145	618	152	1,296

Table 5.2: Distribution of the number of users in each dataset.

5.3.4 Evaluation Method

We evaluate the precision, recall, F1, accuracy, and the PR-AUC for the trained base models. We also use the *latency-weighted F1* (see Section 3.4.2) to determine how well our trained model performs compared to the top-performing models in each shared task.

Lastly, we compute ATTEFA (Amini & Kosseim, 2022) on the test datasets to assess the faithfulness of the user-level attention weights, using Equation 28:

$$\text{ATTEFA} = \frac{1}{N} \sum_{i=1}^N \min \left(\frac{|y_a^i - y_b^i| \cdot \text{JSD}_{max}}{\text{JSD}(\alpha_a^i, \alpha_b^i)}, 1 \right) \quad (28)$$

where y_b^i and y_a^i correspond to the output predictions of the base and the adversarial models, respectively, for user i ; α_b^i and α_a^i represent the user-level attention weights; N is the total number of users in the test dataset; and $JSD_{max} \approx 0.6931$, as stated in Section 5.3.2.

5.4 Results and Discussion

In this section, we present the results of our proposed framework on four mental health detection tasks from the eRisk benchmark: anorexia, depression, pathological gambling, and self-harm. We first evaluate the predictive performance of our base model using standard and latency-weighted metrics, comparing its competitiveness against top-performing systems from the shared tasks. We then analyze the faithfulness of the user-level attention weights via the ATTEFA metric.

5.4.1 Base Model Evaluation

Table 5.3 shows the precision, recall, F1, accuracy, and PR-AUC of the trained based models on each task. The results show that the proposed approach is capable of detecting mental health issues with a high accuracy and a relatively high F1 score.

Dataset	P	R	F1	A	PR-AUC
Anorexia	0.896	0.822	0.857	0.975	0.921
Depression	0.734	0.816	0.773	0.966	0.786
Gambling	0.971	0.981	0.976	0.998	0.998
Self-harm	0.860	0.684	0.762	0.955	0.807

Table 5.3: Base results on test datasets.

To assess the competitiveness of our approach compared to the best-performing models in each eRisk shared task, Table 5.4 provides the ranking of our trained model according to *latency-weighted F1*, compared to the participating models in the corresponding shared task. The results show that our approach is able to provide better or competitive results to the best models and clearly outperforms our Chapter 3 models, showing the validity in our approach.

Rank	Anorexia (55)		Depression (63)		Gambling (50)		Self-harm (56)	
	Team (Run)	lw_{F1}	Team (Run)	lw_{F1}	Team (Run)	lw_{F1}	Team (Run)	lw_{F1}
1	OURS	0.766	NLPGroup-IISERB (0)	0.690	ELiRF-UPV (0)	0.927	UNSL (4)	0.622
2	CLaC (4)	0.690	OURS	0.558	NLP-UNED-2 (3)	0.877	OURS	0.612
3	CLaC (1)	0.690	BLUE (0)	0.540	NLP-UNED-2 (1)	0.876	UNSL (3)	0.583
4	CLaC (3)	0.680	UNSL (2)	0.519	Xabi_EHU (0)	0.875	BLUE (2)	0.578
5	CLaC (2)	0.680	NLPGroup-IISERB (3)	0.511	NLP-UNED-2 (2)	0.875	NLP-UNED (4)	0.564
6	INAOE-CIMAT (3)	0.630	LauSAn (3)	0.498	OURS	0.848	Birmingham (0)	0.551
7	lirimm (0)	0.630	NLPGroup-IISERB (1)	0.496	Xabi_EHU (3)	0.844	NLP-UNED (1)	0.546
8	INAOE-CIMAT (0)	0.620	BioInfo_UAVR (4)	0.494	Xabi_EHU (1)	0.839	NLP-UNED (0)	0.545
9	lirimm (1)	0.620	Sunday-Rocker2 (1)	0.439	NLP-UNED-2 (0)	0.833	BLUE (3)	0.534
10	INAOE-CIMAT (4)	0.610	UNSL (1)	0.426	Xabi_EHU (4)	0.823	NLP-UNED (3)	0.524

Table 5.4: Model rankings for each task based on the *latency-weighted F1* score. The number of participating models in each task is indicated in parentheses in the header row. Records labeled **CLaC** correspond to our models reported in Chapter 3 (see Table 3.3). For details on the corresponding eRisk shared tasks, see the dataset sources listed in Table 5.1.

5.4.2 Faithfulness of the User-level Attention

Table 5.5 shows the ATTEFA scores for each model, along with the average $\mathcal{L}_y$ and $\mathcal{L}_\alpha$ over the test datasets.

Dataset	Initialization	ATTEFA	$\mathcal{L}_y$	$\mathcal{L}_\alpha$
Anorexia	Random	0.527	0.476	0.916
	Trained	0.524	0.481	0.931
Depression	Random	0.497	0.464	0.943
	Trained	0.487	0.455	0.944
Gambling	Random	0.512	0.493	0.968
	Trained	0.731	0.581	0.777
Self-harm	Random	0.538	0.495	0.931
	Trained	0.539	0.490	0.920

Table 5.5: Adversarial training results.

The ATTEFA scores reported in Table 5.5 average around 0.5. In light of the findings by Amini and Kosseim (2022), we infer that, on average, only 10% or fewer of a user’s posts are informative for detecting a mental health problem.⁶ This aligns with prior studies (Amini & Kosseim, 2020; Gui

⁶Amini and Kosseim (2022) reported an ATTEFA score of 0.5 on a synthetic dataset in which only 10 out of 100 tokens per example carried sentiment weights.

et al., 2019; Song, You, Chung, & Park, 2018), which indicate that signs of mental health issues are sparse in social media data. The results also reflect a high degree of faithfulness: if the model attends to a different subset of posts, its predictions deviate considerably from the optimal. This is further supported by the observed $\mathcal{L}_y$ values (approximately 0.5 on average), suggesting that the adversarial models behave similarly to random predictors when attention weights are perturbed.

Another notable observation from Table 5.5 is the high values of $\mathcal{L}_\alpha$ for both types of weight initialization. This indicates that the initialization of the weights does not play a significant role in the convergence, and in almost all cases, maximizing the divergence on the attention weights is a significantly easier objective compared to minimizing the distance between the predictions. The only exception where $\mathcal{L}_\alpha$ does not approach its upper bound can be observed for the Gambling task with the trained base checkpoint as the initialization. We hypothesize that this is due to missing the optimal model checkpoint during adversarial training, as the Gambling dataset is larger than the others (see Table 5.2) and we only saved and evaluated checkpoints at the end of each epoch. We believe that a more frequent checkpointing and evaluation would help bridge the gap on the final value of $\mathcal{L}_\alpha$ between this dataset and the others.

5.5 A Closer Look at the Attention Weights

Amini and Kosseim (2020) (see Chapter 3) demonstrated that attention weights typically correlate with the strength of signals for mental health issues. In their work, they sorted posts in a descending order according to their corresponding attention weights, and obtained the model’s prediction after passing only the top N posts to the model. They observed that increasing the number of posts results in an increasing trend in the precision and a decreasing trend in the recall, indicating that the model usually tends to predict a user as at-risk, if the model only observed the very few highest-weighted posts for that user.

In our study, we perform a variation of their experiment: Instead of starting with only the highest-weighted posts and gradually including the lower-weighted ones, we start by feeding all of the posts to the model and gradually remove the highest-weighted posts one at a time and monitor the prediction flips. Our hypotheses were the following:

- (1) For users predicted as *without risk*, the model should consistently emit a no-risk prediction. This is because the model did not predict those users as *at risk*, even when it observed their highest-weighted posts.
- (2) For users predicted as *at risk*, there should be a cut-off point in the number of posts, at which the model starts considering them as *without risk*. Removing more high-weighted posts should not change the model’s prediction.
- (3) As we expect that, on average, less than 10% of posts contain signs of mental health issues (see Section 5.4.2), the decision flip should often happen after the removal of <10% of the posts.

In our experiment, the first hypothesis is shown to be true for all instances in all 4 datasets. The second hypothesis also turns out to be correct, with the exception of 2 instances in the Depression dataset, where the model’s prediction flips again from *without risk* to *at risk* after removing one extra post after the cut-off. But even in those 2 cases, the model starts predicting *without risk* again after another removal, which remains constant throughout the rest of the process. We believe that this happens due to the model not being fully optimized on the test dataset, resulting in a small number of exceptions where the attention weights do not fully correlate with the degree of the sign of mental health issues. But the fact that this occurs in fewer than 1% of cases highlights the consistency with which the user-level attention weights align with the degree of mental health signal in the posts.

For the third hypothesis, for each test user that was predicted as *at risk*, we calculate the cut-off point at which the decision flip happens. In Table 5.6, we report the average, median, and the standard deviation of the median of the proportion of posts (in percentage) that need to be removed in order to observe a decision flip (from *at risk* to *without risk*). Also in Figure 5.2, we provide the violin plot to visually show the distribution of the cut-off points in proportion to the total number of posts by each user.

Dataset	% Posts		
	Avg	Med	Std
Anorexia	6.27	3.18	7.71
Depression	4.92	1.04	7.99
Gambling	16.03	11.27	16.77
Self-harm	4.61	1.60	6.86

Table 5.6: Average, median, and standard deviation of the percentage of the highest weighted posts per user that need to be removed until a decision flip happens from 1 (*at risk*) to 0 (*without risk*).

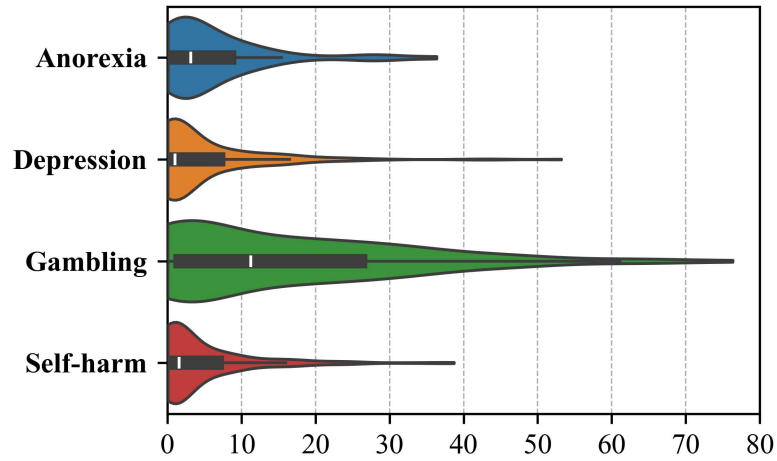


Figure 5.2: Distribution of the percentage of the highest weighted posts per user that need to be removed until a decision flip happens from 1 (*at risk*) to 0 (*without risk*).

As Table 5.6 shows, for the Anorexia, Depression, and Self-harm datasets, the average cut-off is under 10%. The *median + standard deviation* is $\approx 10\%$ for the Anorexia dataset, while being lower than 10% for Depression and Self-harm. These are also observable in the violin plots, where the distribution is mostly centered below 10%. This observation is aligned with the findings of [Amini and Kosseim \(2022\)](#) (see Chapter 4), where they observed that an ATTEFA of ≈ 0.5 occurs when only 10% or fewer samples within a data point are informative to the task.

This behavior, however, is not exhibited in the Gambling dataset, where the average and median cut-off is above 10%. Further analysis is needed to understand why this dataset exhibits this pattern. While this may occur due to the model not being fully optimized on this task (due to the infrequent

checkpointing and evaluation during training, as stated in Section 5.5), it would be useful to perform a deeper study to understand if an inherent characteristic of this dataset has also played a role.

5.6 Limitations

As stated in Section 5.1, the focus of our study was to assess the explainability of the user-level attention mechanism, as opposed to achieving the state of the art. Therefore, we aimed to develop a model with comparable performance to the best performing models in the eRisk shared tasks through minimal hyperparameter tuning and leveraging the test dataset for early stopping and final checkpoint selection. In order to understand the full extent of performance and generalization capability for our approach, the following work remains.

For a fair comparison, the test dataset should remain entirely unseen during training. This also allows for a more accurate assessment of the model’s generalization ability to unseen data. Furthermore, our system should be benchmarked on a wider set of tasks/datasets and against a larger number of systems in the literature. Lastly, we need to further adjust the model or training hyperparameters, and experiment with different encoder models, in order to achieve the most optimal results.

In terms of *explainability assessment*, our current study focuses only on faithfulness and transparency (i.e., what posts were deemed important by the model). In order to fully claim that our system is explainable, we should also assess the notions of plausibility and sufficiency, which are the other two pillars in explainability (Wiegrefe & Pinter, 2019). Such assessments require human assessment, potentially by experts in the field.

Lastly, in order to assess the feasibility of using such a system in a real-life scenario, studies should be done in terms of computational cost and inference latency. In addition, further optimizations are necessary in order to improve the system on those fronts. An additional study would also be necessary to evaluate the practical risks of deploying such a system, in terms of privacy and ethical considerations, and mitigate or reduce such risks as much as possible.

5.7 Conclusion

In this chapter, we introduced a new user-aware, attention-based framework for detecting mental health risks from social media, designed to generate faithful and transparent predictions. Our new architecture modernizes and enhances the approach in Chapter 3 by combining DISORBERT for post-level encoding with a user-level attention mechanism that provides both predictive performance and a potential means for interpretability. To assess the faithfulness of the attention weights, we employed adversarial training and computed ATTEFA (see Chapter 4), a metric designed to quantify how faithfully attention reflects decision-relevant input. Across four tasks in the eRisk shared task series, our system achieved competitive latency-weighted F1 scores and demonstrated high attention faithfulness, particularly in domains where signals are rare and sparsely distributed across user posts.

Our experimental analysis further confirmed that a small fraction of posts (often under 10%) tend to drive the model’s predictions, supporting the hypothesis that mental health signals in social media are both rare and concentrated. These findings underscore the need for transparent and interpretable systems, especially in sensitive domains like mental health, where trust and explainability are essential for practical adoption.

While addressing the limitations presented in Section 5.6 is a prominent future direction, another line of work would be to further move from a system that can detect individual mental health problems to a more general system to assess mental health issues in social media. One promising starting point is to train the model in a multi-task learning fashion (e.g., [Kendall, Gal, and Cipolla, 2018](#)) on a combination of datasets related to the detection of mental health issues, allowing the model to leverage shared knowledge between different tasks.

Chapter 6

Conclusion and Future Work

In this chapter, we conclude the thesis. We first summarize the main contributions with respect to our three research questions. We then discuss limitations and practical considerations, and finally outline directions for future work. In brief, the thesis: 1) developed a strong early risk detection approach in the eRisk setting (Chapter 3); 2) introduced ATTEFA, a quantitative metric to assess attention faithfulness (Chapter 4); and 3) modernized and improved the Chapter 3 approach by integrating DISORBERT with user-level attention and end-to-end fine-tuning, resulting in a user-level, transformer-based system that achieves competitive performance and provides post-level explanations (Chapter 5). Overall, the results show that accurate early detection can be combined with post-level explanations for mental health risks in social media.

6.1 Contributions

In this section, we revisit our research questions (RQ1–RQ3) and summarize the thesis contributions based on the findings presented in Chapters 3–5.

RQ1 (Chapter 3). *Can we build a model that performs strongly on early risk detection of mental health issues on social media?* We developed an effective early detection model on top of the eRisk shared task datasets and showed that a hierarchical attention-based architecture to aggregate posts at the user level is essential for long timelines. Our methodology achieved top results for anorexia detection at eRisk 2019 and provided a platform for later explainability

analyses. This work was published as a chapter in the book *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the eRisk Project* (Amini et al., 2022).

RQ2 (Chapter 4). *Can attention weights provide faithful explanations in text classification tasks, and can their faithfulness be quantified reliably?* We proposed a single-run adversarial training setup that produces a quantitative metric, ATTEFA, to assess attention faithfulness. Across synthetic and real datasets, ATTEFA revealed when single-head attention aligns with the model’s decision logic and when it does not, highlighting that contextual encoders tend to yield lower faithfulness and that faithfulness generally decreases as informative input elements (tokens/posts) become more frequent. This work was published in the proceedings of the BlackBoxNLP 2022 workshop (Amini & Kosseim, 2022)

RQ3 (Chapter 5). *Can we develop an approach that simultaneously achieves competitive early risk detection and produces explanations with demonstrated faithfulness?* We modernized and strengthened the model in Chapter 3 and introduced a hierarchical architecture that integrates a disorder- and social-media-adapted transformer encoder, DISORBERT; (Aragon et al., 2023), with user-level attention. The model achieves competitive latency-weighted F1 on the early detection of depression, anorexia, pathological gambling, and self-harm, while producing post-level explanations whose faithfulness we quantify with ATTEFA. Ablations and perturbation-based analyses show that only a small subset of posts typically drive a positive (at-risk) decision. This work was published in the proceedings of the RANLP 2025 conference (Amini & Kosseim, 2025).

6.2 Limitations

Despite our contributions, our work has several limitations that must be acknowledged.

Data and Generalizability: The datasets used (eRisk datasets based on Reddit posts) are limited in scope and population. They consist of self-selected users who posted in specific sub-reddits related to mental health or other topics, and the ground truth labels (e.g., depression vs. control) are often

based on distant supervision (such as users mentioning a diagnosis). This means that our models are optimized for these benchmark datasets but may not directly generalize to the broader population of social media users or to other platforms like X or Facebook, where writing styles and norms differ. Moreover, the linguistic patterns learned (mostly English data) might not apply to other languages or cultural contexts. Future work could address this by training on more diverse datasets, for example, the recently introduced multilingual mental health corpora (e.g., MENTALRISKES; [Mármol Romero, Moreno-Muñoz, Plaza-Del-Arco, Molina-González, and Montejo-Ráez, 2024](#)) to ensure the approach is not overfit to Reddit-specific language.

Temporal and Real-time Challenges: Our evaluation considered early detection in a simulated setting (with posts revealed in order). In a real deployment, data would stream in real-time and the model might need to handle concept drift (changes in a user’s language over time or slang evolution). We did not explicitly model concept drift or continuously update the model over time. Additionally, the model issues a prediction after a certain number of posts, deciding the optimal moment to raise an alert was simplified in eRisk by the latency metric. In practice, the threshold for alerting (and possibly repeating alerts) is a nuanced issue. Our system could potentially generate false alarms if a user temporarily expresses something that resembles a risk indicator but later context shows otherwise. Conversely, the model might miss slow-building issues if the user’s language changes gradually. These temporal dynamics were not considered and constitute a limitation of our current approach. A possible improvement is to incorporate time gaps between posts into the model or to use reinforcement learning to make the alert emission mechanism more intelligent than our current threshold-based approach.

Scope of Mental Health Conditions: The thesis focused on four conditions for which data was available as part of eRisk datasets. However, in reality, mental health is a broad domain. There are many other disorders (e.g., bipolar disorder, schizophrenia) and nuances (distinguishing between different severities, or detecting changes in state over time) that we did not cover. Each condition can be exhibited differently in language. For instance, the linguistic markers for depression vs. anorexia vs. PTSD can be quite distinct. Our model was able to handle the tasks we tried, but it is unclear how

it would perform on unseen conditions. It might require retraining or additional modules specialized for new conditions. Also, comorbidity (users suffering from multiple conditions) is common in reality but our dataset treats each task separately, which is a limitation. Our current system does not handle comorbidity. Multi-label or multi-class approaches could be needed for a more realistic setting where a single user could be at risk for several issues at once.

Explainability Constraints: Our explainability assessment deliberately focuses on two dimensions: (i) *transparency*, i.e., exposing which posts the model deems important via user-level attention, and (ii) *faithfulness*, i.e., testing whether those highlights are implicated in the decision through our adversarial protocol and ATTEFA. While necessary, these properties are not sufficient for full explainability. A comprehensive study should also assess *plausibility* (whether highlights are convincing and context-appropriate to knowledgeable humans, e.g., clinicians) and *sufficiency* (whether the highlighted subset alone is enough to support the prediction). We did not conduct studies to evaluate neither plausibility nor sufficiency. As a result, our conclusions about explainability are limited: our system provides transparent and quantitatively *faithful* highlights in many cases, but it has not yet been validated for *plausibility* or *sufficiency*.

Ethical and Practical Considerations: There are limitations related to the real-world deployment of such a model that extend beyond technical performance. False positives (flagging someone as at risk when they are not) could cause unnecessary anxiety or privacy invasion. False negatives (missing someone who is at risk) could mean a missed opportunity for help. Our thesis optimized metrics that balance precision and recall, but in practice the acceptable trade-off might differ depending on use (clinical screening vs. personal wellness app, for example). We did not integrate any user feedback loop or human-in-the-loop mechanism, which would likely be necessary in deployment to manage these ethical concerns. Privacy is another limitation: our model was trained on publicly available data, but deploying it might involve monitoring individuals' posts. Any real use would require consent and securing data. While these issues are beyond the scope of our research questions, they are important limitations to acknowledge when considering the impact of this work.

In summary, although our approach marks a step forward in explainable detection of mental

health risks, it is not a complete solution. Addressing the above limitations will be crucial for moving from research to real-world tools. We believe our work provides a foundation that can be extended in more diverse and realistic scenarios.

6.3 Future Work

Building on this thesis, there are several promising avenues for future work.

Clinical Validation and User Studies: One immediate next step is to evaluate the system with mental health professionals and target users. A user study could present clinicians with model outputs and attention-based explanations to see if these align with clinical judgment and whether they aid decision-making. For example, does providing highlighted posts make a clinician more confident in assessing a patient’s condition or identifying relevant symptoms? Similarly, focus groups with users (especially those with lived experience of mental health issues) could reveal how they perceive an AI “assistant” that flags their posts. Such studies would provide insight into the usability and trustworthiness of the explanations. They might also highlight explanation aspects to improve. Perhaps clinicians prefer a combination of highlighted text and a short summary of why it is concerning (which we could generate via Large Language Models (LLMs), as described below). Incorporating human feedback could guide refinement of the explanation mechanism to better meet end-user needs.

Enhanced Explanation Techniques: While we used attention highlighting, future work could explore hybrid explanation approaches. One possible research direction is to generate textual explanations for each decision, in addition to highlighting. For instance, we could use a LLM or fine-tune a (smaller) open-source language model to output sentences like, “The model flagged this user because posts *X*, *Y* and *Z* indicate hopelessness and self-harm ideation.” This could be done by training on a small dataset of explanations or using prompt-based techniques where the model’s own outputs are summarized. Recent work like the ChatGPT-augmented approach of [Belcastro, Cantini, Marozzo, Talia, and Trunfo \(2025\)](#) demonstrates feasibility in transforming technical evidence into human-friendly comments. We would need to ensure these generated explanations remain faithful

(perhaps by constraining them to include only topics present in the highlighted posts). Another explanation enhancement is to incorporate counterfactual reasoning: e.g., “If posts X , Y and Z were not present, the user would likely not be flagged.” This could be derived from our perturbation tests and presented to users for deeper insight. Additionally, integrating other explainable AI methods, such as LIME (Ribeiro et al., 2016) or SHAP (Lundberg & Lee, 2017), with attention could provide multi-faceted explanations; for example, using SHAP to identify which specific words in an important post most affected the prediction, complementing the overall post-level attention.

Expanding to Other Platforms and Modalities: Future research should test and adapt our model for other social media platforms (X, Instagram, TikTok, etc.) and possibly other data modalities. Each platform has different content: X has generally shorter posts, Instagram has images and videos with captions, TikTok primarily has videos with some text. A multi-modal extension of our approach could combine text with image/audio/video analysis. For example, one could incorporate an image recognition module to detect potentially concerning imagery and combine that with text-based analysis. The hierarchical attention concept can extend to modalities by having, for example, an attention mechanism over different sources of data (text vs. image vs. metadata). This is a complex but promising direction, as it could greatly improve coverage of risk signals. Another platform-related challenge is dealing with temporary content (e.g., stories that disappear) or private messages, which raise new data handling issues. But methodologically, adapting to various content types will test the robustness of our attention-based architecture.

Multi-lingual and Cross-cultural Models: As noted in Section 6.2, language use and mental health expression vary across cultures and languages. Future work could train the same model on top of multilingual encoders, or at least test our existing one to non-English data. The attention explanations might also require further adaptations; for example, idioms of sadness in one language might not directly translate. Cross-cultural validation with native speakers or experts would be important to create a more universally applicable tool.

Addressing Broader Mental Health Analytics: Beyond classification of risk, similar techniques could be applied to related problems. For example, through monitoring treatment outcomes or

changes in mental state over time, an explainable model could highlight posts that signify improvement or deterioration in a user’s mood. Another area is content moderation: flagging harmful content (like pro-self-harm posts) for platform intervention. Our attention-based method could provide moderators with quick context on why a post was flagged (e.g., highlight phrases that violate policy). Also, one could adapt the model to detect specific symptoms or factors (like identifying expressions of loneliness vs. anger, which are relevant to different conditions). If combined with a causal analysis, the model might even help in hypothesis generation, like “social isolation language often precedes suicidal ideation posts by weeks”. However that would require careful studies.

Collaboration with Domain Experts: Finally, collaboration with clinicians, psychologists, ethicists, etc. in shaping future work will be essential. This thesis focused on the technical aspects, but deploying such technology involves interdisciplinary considerations. Working closely with practitioners will ensure that improvements are not only technically sound but also clinically meaningful and ethically responsible.

In conclusion, the field of explainable deep learning for mental health has room for further exploration. By extending our current work in the ways outlined above, we can move closer to AI systems that genuinely assist in mental health care (providing early warnings, supporting clinicians, and empowering users), while being transparent and respecting the complexities of mental health. The hope is that such systems, developed and evaluated properly, could one day be integrated into mental health support infrastructure, potentially saving lives and improving well-being through early detection and intervention. We view this thesis as one step along that path.

References

- Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020, November). Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary Perspective. *BMC Medical Informatics and Decision Making*, 20(1).
- Amini, H., Farahnak, F., & Kosseim, L. (2019, June). Natural Language Processing: An Overview. In *Frontiers in Pattern Recognition and Artificial Intelligence* (Vol. 5, pp. 35–55). World Scientific.
- Amini, H., & Kosseim, L. (2020, June). Towards Explainability in Using Deep Learning for the Detection of Anorexia in Social Media. In *Proceedings of the 2020 International Conference on Applications of Natural Language to Information Systems (NLDB 2020)* (pp. 225–235). Online.
- Amini, H., & Kosseim, L. (2022, December). How (Un)Faithful is Attention? In *Proceedings of the 2022 Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackBoxNLP 2022)* (pp. 119–130). Abu Dhabi, UAE.
- Amini, H., & Kosseim, L. (2025, September). From Posts to Predictions: A User-Aware Framework for Faithful and Transparent Detection of Mental Health Risks on Social Media. In *Proceedings of the 2025 International Conference on Recent Advances in Natural Language Processing (RANLP 2025)* (pp. 75–84). Varna, Bulgaria.
- Amini, H., Mohammadi, E., & Kosseim, L. (2022, September). Quick and (maybe not so) Easy Detection of Anorexia in Social Media: To Explainability and Beyond. In *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the eRisk Project* (pp. 141–158). Springer International Publishing.

- Aragon, M., Lopez Monroy, A. P., Gonzalez, L., Losada, D. E., & Montes, M. (2023, July). DisorBERT: A Double Domain Adaptation Model for Detecting Signs of Mental Disorders in Social Media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)* (pp. 15305–15318). Toronto, Canada.
- Aubakirova, M., & Bansal, M. (2016, November). Interpreting Neural Networks to Improve Politeness Comprehension. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)* (pp. 2035–2041). Austin, Texas, USA.
- Bahdanau, D., Cho, K., & Bengio, Y. (2015, May). Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference for Learning Representations (ICLR 2015)*. San Diego, California, USA.
- Baldwin, T., de Marneffe, M. C., Han, B., Kim, Y.-B., Ritter, A., & Xu, W. (2015, July). Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition. In *Proceedings of the 2015 Workshop on Noisy User-generated Text (W-NUT 2015)* (pp. 126–135). Beijing, China.
- Ball, M., Patel, R., Hayes, R. D., Dobson, R. J., & Stewart, R. (2014, November). TextHunter – A User Friendly Tool for Extracting Generic Concepts from Free Text in Clinical Research. In *Proceedings of the 2014 AMIA Annual Symposium* (Vol. 2014). Washington, DC, USA.
- Bauer, B., Norel, R., Leow, A., Rached, Z. A., Wen, B., & Cecchi, G. (2024, May). Using Large Language Models to Understand Suicidality in a Social Media-based Taxonomy of Mental Health Disorders: Linguistic Analysis of Reddit Posts. *JMIR Mental Health*, 11.
- Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961, June). An Inventory for Measuring Depression. *Archives of General Psychiatry*, 4(6), 561–571.
- Belcastro, L., Cantini, R., Marozzo, F., Talia, D., & Trunfio, P. (2025, September). Detecting Mental Disorder on Social Media: A ChatGPT-augmented Explainable Approach. *Online Social Networks and Media*, 48.
- Benton, A., Mitchell, M., & Hovy, D. (2017, April). Multitask Learning for Mental Health Conditions with Limited Social Media Data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)* (pp. 152–162). Valencia, Spain.

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan), 993–1022.
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5, 135–146.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., . . . Amodei, D. (2020, December). Language Models are Few-Shot Learners. In *Proceesings of the 34th Annual Conference on Neural Information Processing Systems (NeurIPS 2020)* (pp. 1877–1901). Online.
- Buddhitha, P., & Inkpen, D. (2023, April). Multi-task Learning to Detect Suicide Ideation and Mental Disorders Among Social Media Users. *Frontiers in Research Metrics and Analytics*, 8.
- Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017, September). Natural Language Processing in Mental Health Applications Using Non-clinical Texts. *Natural Language Engineering*, 23(5), 649–685.
- Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., . . . Raffel, C. (2020, January). Extracting Training Data from Large Language Models. *CoRR*, abs/2012.07805.
- Chancellor, S., Birnbaum, M. L., Caine, K., Silenzio, V. M. B., & De Choudhury, M. (2019, January). A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT* 2019)*. Atlanta, Georgia, USA.
- Chancellor, S., & De Choudhury, M. (2020, March). Methods in predictive techniques for mental health status on ocial media: a critical review. *npj Digital Medicine*, 3(1).
- Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2014, September). One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)* (pp. 2635–2639). Singapore.
- Chen, H., Zheng, G., & Ji, Y. (2020, July). Generating Hierarchical Explanations on Text Classification via Feature Interaction Detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)* (pp. 5578–5593). Online.
- Chen, T., Xu, B., Zhang, C., & Guestrin, C. (2016, April). Training Deep Nets with Sublinear

Memory Cost. *CoRR*, [abs/1604.06174](https://arxiv.org/abs/1604.06174).

- Chim, J., Tsakalidis, A., Gkoumas, D., Atzil-Slonim, D., Ophir, Y., Zirikly, A., . . . Liakata, M. (2024, March). Overview of the CLPsych 2024 Shared Task: Leveraging Large Language Models to Identify Evidence of Suicidality Risk in Online Posts. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)* (pp. 177–190). St. Julians, Malta.
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014, October). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)* (pp. 1724–1734). Doha, Qatar.
- Chrysostomou, G., & Aletras, N. (2021, August). Improving the Faithfulness of Attention-based Explanations with Task-specific Information for Text Classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021) and the 11th International Joint Conference on Natural Language Processing (IJCNLP 2021)* (pp. 477–488). Online.
- Coppersmith, G., Dredze, M., & Harman, C. (2014, June). Quantifying Mental Health Signals in Twitter. In *Proceedings of the 1st Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2014): From Linguistic Signal to Clinical Reality* (pp. 51–60). Baltimore, Maryland, USA.
- Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015, June). CLPsych 2015 Shared Task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2015): From Linguistic Signal to Clinical Reality* (pp. 31–39). Denver, Colorado, USA.
- Davidson, T., Warmusley, D., Macy, M., & Weber, I. (2017, May). Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*. Montreal, Canada.
- De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013, July). Predicting Depression via Social Media. In *Proceedings of the 7th International AAAI Conference on Web and Social Media (ICWSM 2013)* (pp. 128–137). Cambridge, Massachusetts, USA.

- Del Corso, G. M., Gullí, A., & Romani, F. (2005, May). Ranking a Stream of News. In *Proceedings of the 14th International Conference on World Wide Web (WWW 2005)* (pp. 97–106). Chiba, Japan.
- DeVault, D., Georgila, K., Artstein, R., Morbini, F., Traum, D., Scherer, S., ... others (2013, February). Verbal Indicators of Psychological Distress in Interactive Dialogue with a Virtual Human. In *Proceedings of 14th annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2013)* (pp. 193–202). Metz, France.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)* (pp. 4171–4186). Minneapolis, Minnesota, USA.
- DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., & Wallace, B. C. (2020, July). ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)* (pp. 4443–4458). Online.
- Eisenstein, J. (2013, June). What to Do About Bad Language on the Internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)* (pp. 359–369). Atlanta, Georgia, USA.
- Ghaeini, R., Fern, X., & Tadepalli, P. (2018, October–November). Interpreting Recurrent and Attention-Based Neural Models: a Case Study on Natural Language Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)* (pp. 4952–4957). Brussels, Belgium.
- Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G., & Chatzisavvas, K. C. (2017, March). Sentiment Analysis Leveraging Emotions and Word Embeddings. *Expert Systems with Applications*, 69, 214–224.
- Gui, T., Zhang, Q., Zhu, L., Zhou, X., Peng, M., & Huang, X. (2019, October). Depression Detection on Social Media with Reinforcement Learning. In *Proceedings of the 18th China National Conference on Chinese Computational Linguistics (CCL 2019)* (pp. 613–624). Kunming,

China.

- Gullí, A. (2005, May). The Anatomy of a News Search Engine. In *Proceedings of the 14th International Conference on World Wide Web (WWW 2005)* (pp. 880–881). Chiba, Japan.
- Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020, July). Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)* (pp. 8342–8360). Online.
- Hochreiter, S., & Schmidhuber, J. (1997, November). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780.
- Husseini Orabi, A., Buddhitha, P., Husseini Orabi, M., & Inkpen, D. (2018, June). Deep Learning for Depression Detection of Twitter Users. In *Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic (CLPsych 2018)* (pp. 88–97). New Orleans, Louisiana, USA.
- Inkpen, D., Skaik, R., Buddhitha, P., Angelov, D., & Fredenburgh, M. T. (2021, September). uOttawa at eRisk 2021: Automatic Filling of the Beck’s Depression Inventory Questionnaire using Deep Learning. In *Working Notes of CLEF 2021 – Conference and Labs of the Evaluation Forum*. Online.
- Ishihara, S. (2023, July). Training Data Extraction From Pre-trained Language Models: A Survey. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)* (pp. 260–275). Toronto, Canada.
- Jackson, R. G., Patel, R., Jayatilleke, N., Kolliakou, A., Ball, M., Gorrell, G., ... Stewart, R. (2017, January). Natural Language Processing to Extract Symptoms of Severe Mental Illness from Clinical Text: The Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) Project. *BMJ open*, 7(1).
- Jacovi, A., & Goldberg, Y. (2020, July). Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)* (pp. 4198–4205). Online.
- Jain, S., & Wallace, B. C. (2019, June). Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:*

- Human Language Technologies (NAACL-HLT 2019)* (pp. 3543–3556). Minneapolis, Minnesota, USA.
- Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., . . . Mark, R. G. (2016, May). MIMIC-III, a Freely Accessible Critical Care Database. *Scientific Data*, 3(1).
- Karlekar, S., Niu, T., & Bansal, M. (2018, June). Detecting Linguistic Characteristics of Alzheimer’s Dementia by Interpreting Neural Models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)* (pp. 701–707). New Orleans, Louisiana, USA.
- Kendall, A., Gal, Y., & Cipolla, R. (2018, June). Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In *Proceedings of the 2018 Conference on Computer Vision and Pattern Recognition (CVPR 2018)*. Salt Lake City, Utah, USA.
- Kingma, D. P., & Ba, J. (2015, May). Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference for Learning Representations (ICLR 2015)*. San Diego, California, USA.
- Koh, P. W., & Liang, P. (2017, August). Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 2017 International Conference on Machine Learning (ICML 2017)* (p. 1885–1894). Sydney, Australia.
- Lang, K. (1995, July). NewsWeeder: Learning to Filter Netnews. In *Proceedings of the 12th International Conference on Machine Learning (ICML 1995)* (pp. 331–339). Tahoe City, California, USA.
- LeCun, Y., Haffner, P., Bottou, L., & Bengio, Y. (1999, October). Object Recognition with Gradient-based Learning. In *Shape, Contour and Grouping in Computer Vision* (pp. 319–345). Springer Berlin Heidelberg.
- Lee, J., Shin, J.-H., & Kim, J.-S. (2017, September). Interactive Visualization and Manipulation of Attention-based Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017): System Demonstrations* (pp. 121–126). Copenhagen, Denmark.
- Lin, Z., Feng, M., Nogueira dos Santos, C., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017, April). A Structured Self-Attentive Sentence Embedding. In *Proceedings of the 5th International*

- Conference on Learning Representations (ICLR 2017)*. Toulon, France.
- Lipton, Z. C. (2016, June). The mythos of model interpretability. *CoRR*, *abs/1606.03490*.
- Losada, D. E., & Crestani, F. (2016, September). A Test Collection for Research on Depression and Language Use. In *Proceedings of CLEF 2016 – Conference and Labs of the Evaluation Forum*. Évora, Portugal.
- Losada, D. E., Crestani, F., & Parapar, J. (2017, September). CLEF 2017 eRisk Overview: Early Risk Prediction on the Internet: Experimental Foundations. In *Working Notes of CLEF 2017 – Conference and Labs of the Evaluation Forum*. Dublin, Ireland.
- Losada, D. E., Crestani, F., & Parapar, J. (2018, September). Overview of eRisk 2018: Early Risk Prediction on the Internet (Extended Lab Overview). In *Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum*. Avignon, France.
- Losada, D. E., Crestani, F., & Parapar, J. (2019, September). Overview of eRisk 2019: Early Risk Prediction on the Internet (Extended Lab Overview). In *Working Notes of CLEF 2019 – Conference and Labs of the Evaluation Forum*. Lugano, Switzerland.
- Losada, D. E., Crestani, F., & Parapar, J. (2020, September). Overview of eRisk 2020: Early Risk Prediction on the Internet (Extended Lab Overview). In *Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum*. Thessaloniki, Greece.
- Loshchilov, I., & Hutter, F. (2019, May). Decoupled Weight Decay Regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*. New Orleans, Louisiana, USA.
- Lundberg, S. M., & Lee, S.-I. (2017, December). A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach, California, USA.
- Luong, T., Pham, H., & Manning, C. D. (2015, September). Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)* (pp. 1412–1421). Lisbon, Portugal.
- Lynn, V., Goodman, A., Niederhoffer, K., Loveys, K., Resnik, P., & Schwartz, H. A. (2018, June). CLPsych 2018 Shared Task: Predicting Current and Future Psychological Health from Childhood Essays. In *Proceedings of the 5th Workshop on Computational Linguistics and*

- Clinical Psychology: From Keyboard to Clinic (CLPsych 2018)* (pp. 37–46). New Orleans, Louisiana, USA.
- Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)* (pp. 142–150). Portland, Oregon, USA.
- Mármol Romero, A. M., Moreno-Muñoz, A., Plaza-Del-Arco, F. M., Molina-González, M. D., & Montejo-Ráez, A. (2024, May). MentalRiskES: A New Corpus for Early Detection of Mental Disorders in Spanish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 11204–11214). Turin, Italy.
- Maupomé, D., Queudot, M., & Meurs, M.-J. (2019, May). Inter and Intra Document Attention for Depression Risk Assessment. In *Proceedings of the 2019 Canadian Conference on Artificial Intelligence (Canadian AI 2019)* (pp. 333–341). Kingston, Canada.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013, May). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the 2013 International Conference on Learning Representations (ICLR 2013)*. Scottsdale, Arizona, USA.
- Milne, D. N., Pink, G., Hachey, B., & Calvo, R. A. (2016, June). CLPsych 2016 Shared Task: Triaging content in online peer-support forums. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2016)* (pp. 118–127). San Diego, California, USA.
- Mohammadi, E., Amini, H., & Kosseim, L. (2019a, June). CLaC at CLPsych 2019: Fusion of Neural Features and Predicted Class Probabilities for Suicide Risk Assessment Based on Online Posts. In *Proceedings of the 6th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2019)* (pp. 34–38). Minneapolis, Minnesota, USA.
- Mohammadi, E., Amini, H., & Kosseim, L. (2019b, September). Quick and (maybe not so) Easy Detection of Anorexia in Social Media Posts. In *Working Notes of CLEF 2019 – Conference and Labs of the Evaluation Forum*. Lugano, Switzerland.
- Montejo-Ráez, A., Molina-González, M. D., Jiménez-Zafra, S. M., Ángel García-Cumbreras, M., &

- García-López, L. J. (2024, August). A Survey on Detecting Mental Disorders with Natural Language Processing: Literature Review, Trends and Challenges. *Computer Science Review*, 53.
- Moradi, P., Kambhatla, N., & Sarkar, A. (2021, April). Measuring and Improving Faithfulness of Attention in Neural Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021)* (pp. 2791–2802). Online.
- Mullenbach, J., Wiegrefe, S., Duke, J., Sun, J., & Eisenstein, J. (2018, June). Explainable Prediction of Medical Codes from Clinical Text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)* (pp. 1101–1111). New Orleans, Louisiana, USA.
- Murarka, A., Radhakrishnan, B., & Ravichandran, S. (2021, April). Classification of Mental Illnesses on Social Media Using RoBERTa. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis (LOUHI 2021)* (pp. 59–68). online.
- National Institute of Mental Health. (2025). *Mental Illness*. <https://www.nimh.nih.gov/health/statistics/mental-illness>. (Accessed: 2025-08-31)
- Ofoghi, B., Mann, M., & Verspoor, K. (2016, January). Towards Early Discovery of Salient Health Threats: A Social Media Emotion Classification Technique. In *Proceedings of the 2016 Pacific Symposium on Biocomputing (PSB 2016)* (pp. 504–515). Kohala Coast, Hawaii, USA.
- Parapar, J., Martín-Rodilla, P., Losada, D. E., & Crestani, F. (2021, September). Overview of eRisk at CLEF 2021: Early Risk Prediction on the Internet (Extended Overview). In *Working Notes of CLEF 2021 – Conference and Labs of the Evaluation Forum*. Bucharest, Romania.
- Parapar, J., Martín-Rodilla, P., Losada, D. E., & Crestani, F. (2022, September). Overview of eRisk at CLEF 2022: Early Risk Prediction on the Internet (Extended Overview). In *Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum*. Bologna, Italy.
- Parapar, J., Martín-Rodilla, P., Losada, D. E., & Crestani, F. (2023, September). Overview of eRisk at CLEF 2023: Early Risk Prediction on the Internet (Extended Overview). In *Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum*. Thessaloniki, Greece.
- Parapar, J., Martín-Rodilla, P., Losada, D. E., & Crestani, F. (2024, September). Overview of eRisk

- at CLEF 2024: Early Risk Prediction on the Internet (Extended Overview). In *Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum*. Grenoble, France.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019, December). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proceedings of the 2019 Conference on Neural Information Processing Systems (NeurIPS 2019)*. Vancouver, Canada.
- Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., ... others (2018, October). The Lancet Commission on Global Mental Health and Sustainable Development. *The Lancet*, 392(10157), 1553–1598.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011, November). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Pennington, J., Socher, R., & Manning, C. (2014, October). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)* (pp. 1532–1543). Doha, Qatar.
- Pestian, J., Nasrallah, H., Matykiewicz, P., Bennett, A., & Leenaars, A. (2010, August). Suicide Note Classification Using Natural Language Processing: A Content Analysis. *Biomedical Informatics Insights*, 3, 19–28.
- Pruthi, D., Gupta, M., Dhingra, B., Neubig, G., & Lipton, Z. C. (2020, July). Learning to Deceive with Attention-Based Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)* (pp. 4782–4793). Online.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016)* (pp. 1135–1144). San Diego, California, USA.
- Rudin, C. (2019, May). Stop Explaining Black Box Machine Learning Models for High-stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1(5), 206–215.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017, October). Grad-cam: Visual Explanations from Deep Networks via Gradient-based Localization. In

- Proceedings of the 2017 IEEE international conference on computer vision (ICCV 2017)* (pp. 618–626). Venice, Italy.
- Serrano, S., & Smith, N. A. (2019, July). Is Attention Interpretable? In *Proceedings of 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)* (pp. 2931–2951). Florence, Italy.
- Shapley, L. S. (1953, December). A Value for n-Person Games. In *Contributions to the Theory of Games II* (pp. 307–317). Princeton University Press.
- Shen, J. H., & Rudzicz, F. (2017, August). Detecting Anxiety Through Reddit. In *Proceedings of the 4th Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (CLPsych 2017)* (pp. 58–65). Vancouver, Canada.
- Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership Inference Attacks Against Machine Learning Models. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP 2017)* (pp. 3–18). Los Alamitos, CA, USA.
- Skaik, R., & Inkpen, D. (2020, December). Using Social Media for Mental Health Surveillance: A Review. *ACM Computing Surveys*, 53(6), 1–31.
- Skaik, R., & Inkpen, D. (2022, September). Predicting Depression in Canada by Automatic Filling of Beck’s Depression Inventory Questionnaire. *IEEE Access*, 10, 102033-102047.
- Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013, October). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)* (pp. 1631–1642). Seattle, Washington, USA.
- Song, H., You, J., Chung, J.-W., & Park, J. C. (2018, December). Feature Attention Network: Interpretable Depression Detection from Social Media. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation (PACLIC 32)* (pp. 613–622). Hong Kong.
- Struik, L. L., & Baskerville, N. B. (2014, July). The role of Facebook in Crush the Crave, a Mobile-and Social Media-based Smoking Cessation Intervention: Qualitative Framework Analysis of Posts. *Journal of Medical Internet Research*, 16(7).
- Sundararajan, M., Taly, A., & Yan, Q. (2017, August). Axiomatic Attribution for Deep Networks.

- In *Proceedings of the 2017 International Conference on Machine Learning (ICML 2017)* (pp. 3319–3328). Sydney, Australia.
- Thompson, J. J., Leung, B. H., Blair, M. R., & Taboada, M. (2017, December). Sentiment Analysis of Player Chat Messaging in the Video Game StarCraft 2: Extending a Lexicon-based Bodel. *Knowledge-based Systems, 137*, 149–162.
- Thornicroft, G., Chatterji, S., Evans-Lacko, S., Gruber, M., Sampson, N., Aguilar-Gaxiola, S., . . . others (2017, February). Undertreatment of People with Major Depressive Disorder in 21 Countries. *The British Journal of Psychiatry, 210*(2), 119–124.
- Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019, August). What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. In *Proceedings of the 4th machine learning for healthcare conference* (pp. 359–380). Ann Arbor, Michigan, USA.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., . . . Scialom, T. (2023, July). Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR, 2307.09288*.
- Trotzek, M., Koitka, S., & Friedrich, C. M. (2018, September). Word Embeddings and Linguistic Metadata at the CLEF 2018 Tasks for Early Detection of Depression and Anorexia. In *Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum*. Avignon, France.
- Tsakalidis, A., Chim, J., Bilal, I. M., Zirikly, A., Atzil-Slonim, D., Nanni, F., . . . Liakata, M. (2022, July). Overview of the CLPsych 2022 Shared Task: Capturing Moments of Change in Longitudinal User Posts. In *Proceedings of the 8th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2022)* (pp. 184–198). Seattle, Washington, USA.
- Tseriotou, T., Chim, J., Klein, A., Shamir, A., Dvir, G., Ali, I., . . . Liakata, M. (2025, May). Overview of the CLPsych 2025 Shared Task: Capturing Mental Health Dynamics from Social Media Timelines. In *Proceedings of the 10th workshop on computational linguistics and clinical psychology (clpsych 2025)* (pp. 193–217). Albuquerque, New Mexico, USA.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017, December). Attention is All you Need. In *Proceesings of the 31st Annual Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach, California, USA.
- Wang, B., Zhao, Y., Lu, X., & Qin, B. (2023, January). Cognitive Distortion Based Explainable

- Depression Detection and Analysis Technologies for the Adolescent Internet Users on Social Media. *Frontiers in Public Health*, 10.
- Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2016, November). Attention-based LSTM for Aspect-level Sentiment Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)* (pp. 606–615). Austin, Texas, USA.
- Wang, Y., Inkpen, D., & Kirinde Gamaarachchige, P. (2024, March). Explainable Depression Detection Using Large Language Models on Social Media Data. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)* (pp. 108–126). St. Julians, Malta.
- Wiegrefe, S., & Pinter, Y. (2019, November). Attention is not not Explanation. In *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. 11–20). Hong Kong, China.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., . . . Rush, A. M. (2020, October). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): System Demonstrations* (pp. 38–45). Online.
- World Health Organization. (2022a). *Depression: Key Facts*. <https://www.who.int/news-room/fact-sheets/detail/depression>. (Accessed: 2025-08-31)
- World Health Organization. (2022b). *Mental Disorders*. <https://www.who.int/news-room/fact-sheets/detail/mental-disorders>. (Accessed: 2025-08-31)
- World Health Organization. (2022c). *World Mental Health Report: Transforming Mental Health for All*. <https://www.who.int/publications/i/item/9789240049338>. (Accessed: 2025-08-31)
- Xie, Q., Ma, X., Dai, Z., & Hovy, E. (2017, July). An Interpretable Knowledge Transfer Model for Knowledge Base Completion. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)* (pp. 950–962). Vancouver, Canada.
- Yang, C. C., Jiang, L., Yang, H., & Tang, X. (2012, August). Detecting Signals of Adverse Drug Reactions from Health Consumer Contributed Content in Social Media. In *Proceedings of the*

- 2021 ACM SIGKDD Workshop on Health Informatics (HI-KDD 2012). Beijing, China.
- Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016, June). Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)* (pp. 1480–1489). San Diego, California, USA.
- Yates, A., Cohan, A., & Goharian, N. (2017, September). Depression and Self-Harm Risk Assessment in Online Forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)* (pp. 2968–2978). Copenhagen, Denmark.
- Yin, J., Karimi, S., Lampert, A., Cameron, M., Robinson, B., & Power, R. (2015, July). Using Social Media to Enhance Emergency Situation Awareness. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*. Buenos Aires, Argentina.
- Zhuang, L., Wayne, L., Ya, S., & Jun, Z. (2021, August). A Robustly Optimized BERT Pre-training Approach with Post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics (CCL 2021)* (pp. 1218–1227). Huhhot, China.
- Zintgraf, L. M., Cohen, T. S., Adel, T., & Welling, M. (2017, April). Visualizing Deep Neural Network Decisions: Prediction Difference Analysis. In *Proceedings of the 2017 International Conference on Learning Representations (ICLR 2017)*. Toulon, France.
- Zirikly, A., Resnik, P., Uzuner, Ö., & Hollingshead, K. (2019, June). CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts. In *Proceedings of the 6th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2019)* (pp. 24–33). Minneapolis, Minnesota, USA.
- Zogan, H., Razzak, I., Wang, X., Jameel, S., & Xu, G. (2022, January). Explainable Depression Detection with Multi-aspect Features Using a Hybrid Deep Learning Model on Social Media. *World Wide Web*, 25(1), 281–304.

Appendix A

Synthetic Datasets and Adversarial Results

This appendix complements Chapter 4 by providing additional information on the ATTEFA experiments.

A.1 Samples from the *Mock*-* Datasets

Tables A.1, A.2, A.3, A.4, A.5, A.6, A.7, and A.8 contain 60-token positive and negative examples for each Mock-* dataset. The samples have 1, 2, 5, 10, 15, 30, 45, and 60 informative tokens in the Mock-1, Mock-2, Mock-5, Mock-10, Mock-1q, Mock-2q, Mock-3q, and Mock-4q datasets, respectively. These tokens are shown in colors blue or red, depending on their positive or negative sentiment polarity. To show the frequency of informative tokens in each sample, these tokens are accompanied by superscripted numbers in purple, which range from 1 to the number of informative tokens in each sample.

Positive and negative tokens were selected randomly from the following gazetteers, respectively:

- <https://ptrckprry.com/course/ssd/data/positive-words.txt>
- <https://ptrckprry.com/course/ssd/data/negative-words.txt>

[illegible]

Table A.1: Positive and negative sample from the Mock-1 dataset.

Positive	Negative
something something something some- thing something something something something something something some- thing something something something something something eagerly ¹ some- thing something something something something something something some- thing something something something something something gaily ² something something thrive ³ something something something something something some- thing resolute ⁴ upliftment ⁵ something something something something some- thing something something something something something something some- thing something something something something something something	something something something some- thing something something something something something something some- thing something violation ¹ something something something something un- realistic ² something something dents ³ something something something worry ⁴ something something something some- thing something something something something something something some- thing something something something something something something wom- anizing ⁵ something something some- thing something something something something something something some- thing something something something something something something some- thing

Table A.3: Positive and negative sample from the Mock-5 dataset.

Positive	Negative
something something something re- funded ¹ something something some- thing something something something something something something rejuve- nated ² something something eagerly ³ something something something some- thing something something something something something something some- thing something something gaily ⁴ some- thing stronger ⁵ thrive ⁶ something re- joicingly ⁷ something something some- thing something resolute ⁸ upliftment ⁹ something something amity ¹⁰ something something something something some- thing something something something something something something some- thing something something something	something messed ¹ something some- thing something something something something something something some- thing something violation ² something something something something un- realistic ³ something something dents ⁴ something something something worry ⁵ something something something some- thing something something something something something something some- thing something something something raped ⁶ something something womaniz- ing ⁷ something something something something something illusory ⁸ some- thing something something nervous- ness ⁹ buzzing ¹⁰ something something something something something some- thing

Table A.4: Positive and negative sample from the Mock-10 dataset.

Positive	Negative
something something something re- funded ¹ something something some- thing reconciliation ² something some- thing something wonder ³ something rejuvenated ⁴ something something ea- gerly ⁵ something something some- thing inestimably ⁶ something some- thing something something something something something something some- thing gaily ⁷ something stronger ⁸ thrive ⁹ something rejoicingly ¹⁰ unbeatable ¹¹ something something something res- olute ¹² upliftment ¹³ something some- thing amity ¹⁴ softer ¹⁵ something some- thing something something something something something something some- thing something something something something something	something messed ¹ something some- thing suspicions ² something uproarious ³ something something something some- thing something violation ⁴ something something something something un- realistic ⁵ something something dents ⁶ something something something worry ⁷ something something pan ⁸ something something something something some- thing something something something something something something raped ⁹ something something womanizing ¹⁰ something something something screw- up ¹¹ something illusory ¹² something something something nervousness ¹³ buzzing ¹⁴ something something radi- cals ¹⁵ something something something

Table A.5: Positive and negative sample from the Mock-1q dataset.

Positive	Negative
exaltation¹ something something re- funded² fortuitously³ something afford⁴ reconciliation⁵ something something something wonder⁶ something rejuven- ated⁷ something something eagerly⁸ mesmerized⁹ something exhilarating¹⁰ inestimably¹¹ something smilingly¹² something something something some- thing priceless¹³ something something gaily¹⁴ something stronger¹⁵ thrive¹⁶ superior¹⁷ rejoicingly¹⁸ unbeatable¹⁹ something something something reso- lute²⁰ upliftment²¹ something irreplace- able²² amity²³ softer²⁴ something first- class²⁵ something appreciated²⁶ coura- geously²⁷ something outperformed²⁸ il- luminati²⁹ something something some- thing something willing³⁰ something	something messed¹ something some- thing suspicions² something uproarious³ something something something some- thing something violation⁴ inordinate⁵ something something picketed⁶ un- realistic⁷ cracked⁸ something dents⁹ something wretch¹⁰ something worry¹¹ flairs¹² importune¹³ pan¹⁴ something something bland¹⁵ something accusa- tions¹⁶ something something some- thing loveless¹⁷ something backward¹⁸ raped¹⁹ something something womaniz- ing²⁰ petrified²¹ shamefulness²² scandal- ized²³ screw-up²⁴ something illusory²⁵ something stumble²⁶ something ner- vousness²⁷ buzzing²⁸ something some- thing radicals²⁹ untouched³⁰ something something

Table A.6: Positive and negative sample from the Mock-2q dataset.

Positive	Negative
exaltation ¹ like ² educated ³ refunded ⁴ fortuitously ⁵ something afford ⁶ reconcil- iation ⁷ sporty ⁸ everlasting ⁹ something wonder ¹⁰ outperforms ¹¹ rejuvenated ¹² something something eagerly ¹³ mesmer- ized ¹⁴ something exhilarating ¹⁵ ines- timably ¹⁶ something smilingly ¹⁷ law- ful ¹⁸ something something something priceless ¹⁹ enlighten ²⁰ adventurous ²¹ gaily ²² something stronger ²³ thrive ²⁴ superior ²⁵ rejoicingly ²⁶ unbeatable ²⁷ something something recommended ²⁸ resolute ²⁹ upliftment ³⁰ sufficiently ³¹ ir- replaceable ³² amity ³³ softer ³⁴ shimmer- ing ³⁵ first-class ³⁶ something appreci- ated ³⁷ courageously ³⁸ boundless ³⁹ out- performed ⁴⁰ illuminati ⁴¹ something vir- tuously ⁴² savvy ⁴³ brand-new ⁴⁴ willing ⁴⁵ something	disinterested ¹ messed ² something distortion ³ suspicions ⁴ hum ⁵ uproarious ⁶ something something desert ⁷ calamity ⁸ unlawfully ⁹ violation ¹⁰ inordinate ¹¹ something invalidity ¹² picketed ¹³ unrealistic ¹⁴ cracked ¹⁵ feverish ¹⁶ dents ¹⁷ something wretch ¹⁸ strangest ¹⁹ worry ²⁰ flairs ²¹ importune ²² pan ²³ something something bland ²⁴ fanati- cal ²⁵ accusations ²⁶ something stigma ²⁷ improbable ²⁸ loveless ²⁹ misman- age ³⁰ backward ³¹ raped ³² something something womanizing ³³ petrified ³⁴ shamefulness ³⁵ scandalized ³⁶ screw- up ³⁷ something illusory ³⁸ symptoms ³⁹ stumble ⁴⁰ something nervousness ⁴¹ buzzing ⁴² something upset ⁴³ radicals ⁴⁴ untouched ⁴⁵ something something

Table A.7: Positive and negative sample from the Mock-3q dataset.

Positive	Negative
exaltation¹ like² educated³ refunded⁴ fortuitously⁵ immaculate⁶ afford⁷ reconciliation⁸ sporty⁹ everlasting¹⁰ quiet¹¹ wonder¹² outperforms¹³ reju- venated¹⁴ astonishment¹⁵ exaltingly¹⁶ eagerly¹⁷ mesmerized¹⁸ flawless¹⁹ exhilarating²⁰ inestimably²¹ roomier²² smilingly²³ lawful²⁴ trustworthiness²⁵ reconcile²⁶ marvelous²⁷ priceless²⁸ enlighten²⁹ adventurous³⁰ gaily³¹ fabulous³² stronger³³ thrive³⁴ superior³⁵ rejoicingly³⁶ unbeatable³⁷ redemption³⁸ lean³⁹ recommended⁴⁰ resolute⁴¹ upliftment⁴² sufficiently⁴³ irreplace- able⁴⁴ amity⁴⁵ softer⁴⁶ shimmering⁴⁷ first-class⁴⁸ fervor⁴⁹ appreciated⁵⁰ courageously⁵¹ boundless⁵² outper- formed⁵³ illuminati⁵⁴ painlessly⁵⁵ virtuously⁵⁶ savvy⁵⁷ brand-new⁵⁸ willing⁵⁹ easygoing⁶⁰	disinterested¹ messed² impugn³ distor- tion⁴ suspicions⁵ hum⁶ uproarious⁷ re- volting⁸ clunky⁹ desert¹⁰ calamity¹¹ un- lawfully¹² violation¹³ inordinate¹⁴ re- volt¹⁵ invalidity¹⁶ picketed¹⁷ unrealis- tic¹⁸ cracked¹⁹ feverish²⁰ dents²¹ hyp- ocrite²² wretch²³ strangest²⁴ worry²⁵ flairs²⁶ importune²⁷ pan²⁸ distress- ingly²⁹ apathetically³⁰ bland³¹ fanati- cal³² accusations³³ grudge³⁴ stigma³⁵ improbable³⁶ loveless³⁷ mismanage³⁸ backward³⁹ raped⁴⁰ feint⁴¹ antagonis- tic⁴² womanizing⁴³ petrified⁴⁴ shame- fulness⁴⁵ scandalized⁴⁶ screw-up⁴⁷ en- trap⁴⁸ illusory⁴⁹ symptoms⁵⁰ stumble⁵¹ incompetent⁵² nervousness⁵³ buzzing⁵⁴ wince⁵⁵ upset⁵⁶ radicals⁵⁷ untouched⁵⁸ draconian⁵⁹ decayed⁶⁰

Table A.8: Positive and negative sample from the Mock-4q dataset.

A.2 All Results from the Adversarial Setup

Tables [A.9](#) and [A.10](#) constitute extended versions of Table [4.3](#), which include the results from the adversarial setup for each individual random seed that was used in the training of the adversarial models.

Dataset	Seed	LSTM					MLP				
		Epoch	F1	TVD	JSD	ATTEFa	Epoch	F1	TVD	JSD	ATTEFa
Mock-1	10	7	0.923	0.025	0.693	0.050	47	0.662	0.222	0.393	0.784
	50	16	0.971	0.004	0.693	0.008	1	0.000	0.244	0.393	0.860
	257	15	0.943	0.017	0.693	0.033	1	0.000	0.219	0.393	0.774
	500231	20	0.960	0.011	0.693	0.021	76	0.662	0.223	0.393	0.787
	100078	13	0.968	0.006	0.693	0.011	9	0.000	0.247	0.393	0.872
	12504	7	0.916	0.029	0.693	0.057	11	0.000	0.249	0.393	0.879
	90754789	16	0.963	0.008	0.693	0.015	7	0.000	0.218	0.393	0.767
	8988812	3	0.926	0.026	0.693	0.052	30	0.662	0.223	0.393	0.785
	2	9	0.952	0.012	0.693	0.025	2	0.000	0.235	0.393	0.828
Mock-2	10	2	0.982	0.015	0.692	0.030	1	0.662	0.243	0.670	0.504
	50	4	0.977	0.016	0.693	0.032	1	0.000	0.248	0.670	0.513
	257	12	0.978	0.016	0.693	0.032	1	0.000	0.244	0.670	0.506
	500231	5	0.964	0.024	0.693	0.047	2	0.662	0.242	0.670	0.502
	100078	15	0.975	0.016	0.693	0.033	1	0.000	0.247	0.670	0.510
	12504	19	0.960	0.024	0.693	0.048	1	0.000	0.250	0.670	0.519
	90754789	3	0.978	0.017	0.693	0.034	1	0.000	0.242	0.670	0.502
	8988812	4	0.984	0.012	0.693	0.024	1	0.662	0.243	0.670	0.504
	2	19	0.997	0.007	0.693	0.015	1	0.000	0.250	0.670	0.519
Mock-5	10	12	1.000	0.002	0.693	0.003	1	0.662	0.249	0.686	0.504
	50	3	1.000	0.002	0.693	0.003	17	0.000	0.246	0.686	0.497
	257	2	0.998	0.003	0.693	0.006	4	0.000	0.246	0.686	0.497
	500231	20	1.000	0.002	0.693	0.003	2	0.000	0.247	0.686	0.499
	100078	5	0.999	0.002	0.693	0.004	16	0.000	0.246	0.686	0.497
	12504	7	1.000	0.002	0.693	0.003	2	0.000	0.246	0.686	0.497
	90754789	3	1.000	0.002	0.693	0.003	34	0.000	0.246	0.686	0.497
	8988812	5	1.000	0.002	0.693	0.003	1	0.662	0.249	0.686	0.504
	2	7	1.000	0.002	0.693	0.003	2	0.000	0.246	0.686	0.497
Mock-10	10	21	1.000	0.001	0.693	0.001	1	0.662	0.251	0.686	0.508
	50	25	1.000	0.001	0.693	0.001	80	0.000	0.248	0.686	0.501
	257	6	1.000	0.001	0.693	0.001	13	0.000	0.248	0.686	0.501
	500231	8	1.000	0.001	0.693	0.001	2	0.000	0.249	0.686	0.503
	100078	16	1.000	0.001	0.693	0.001	80	0.000	0.248	0.686	0.501
	12504	4	1.000	0.001	0.693	0.001	6	0.000	0.248	0.686	0.501
	90754789	9	1.000	0.001	0.693	0.001	24	0.000	0.248	0.686	0.501
	8988812	28	1.000	0.001	0.693	0.001	1	0.662	0.251	0.686	0.508
	2	12	1.000	0.001	0.693	0.001	4	0.000	0.248	0.686	0.501
Mock-1q	10	20	1.000	0.000	0.693	0.001	70	0.000	0.247	0.690	0.497
	50	40	1.000	0.000	0.693	0.001	10	1.000	0.001	0.688	0.002
	257	39	1.000	0.000	0.693	0.001	79	0.093	0.235	0.690	0.473
	500231	13	1.000	0.000	0.693	0.001	80	0.093	0.235	0.690	0.473
	100078	17	1.000	0.000	0.693	0.001	6	1.000	0.002	0.688	0.003
	12504	14	1.000	0.000	0.693	0.001	6	1.000	0.002	0.689	0.003
	90754789	12	1.000	0.000	0.693	0.001	44	0.000	0.247	0.690	0.497
	8988812	15	1.000	0.000	0.693	0.001	28	0.000	0.247	0.690	0.497
	2	39	1.000	0.000	0.693	0.001	6	1.000	0.002	0.689	0.004
Mock-2q	10	25	1.000	0.000	0.678	0.001	80	0.305	0.203	0.690	0.408
	50	80	1.000	0.000	0.681	0.000	79	0.700	0.212	0.690	0.426
	257	79	1.000	0.000	0.681	0.000	4	0.997	0.002	0.691	0.003
	500231	5	0.998	0.002	0.678	0.003	40	0.995	0.003	0.691	0.005
	100078	23	0.999	0.001	0.681	0.001	80	0.089	0.236	0.690	0.474
	12504	8	1.000	0.000	0.682	0.000	79	0.683	0.230	0.690	0.461
	90754789	12	1.000	0.000	0.681	0.000	4	0.998	0.001	0.691	0.002
	8988812	3	1.000	0.000	0.670	0.000	10	0.997	0.002	0.691	0.004
	2	79	1.000	0.000	0.670	0.000	6	0.999	0.001	0.690	0.001
Mock-3q	10	5	1.000	0.000	0.674	0.000	5	0.998	0.001	0.691	0.002
	50	16	1.000	0.000	0.675	0.000	4	0.997	0.002	0.691	0.003
	257	9	1.000	0.000	0.674	0.001	2	1.000	0.000	0.690	0.000
	500231	12	1.000	0.000	0.691	0.000	44	1.000	0.000	0.691	0.000
	100078	10	1.000	0.000	0.691	0.000	13	1.000	0.000	0.691	0.000
	12504	4	1.000	0.000	0.689	0.000	5	0.999	0.001	0.691	0.001
	90754789	46	1.000	0.000	0.675	0.000	11	1.000	0.000	0.691	0.000
	8988812	71	1.000	0.000	0.686	0.000	25	0.999	0.001	0.691	0.001
	2	12	1.000	0.000	0.674	0.000	5	0.998	0.001	0.691	0.002
Mock-4q	10	6	1.000	0.000	0.683	0.000	3	1.000	0.000	0.690	0.000
	50	7	1.000	0.000	0.683	0.000	3	1.000	0.000	0.690	0.000
	257	5	1.000	0.000	0.683	0.000	3	1.000	0.000	0.690	0.000
	500231	12	1.000	0.000	0.683	0.000	2	1.000	0.000	0.690	0.000
	100078	9	1.000	0.000	0.683	0.000	3	1.000	0.000	0.690	0.000
	12504	16	1.000	0.000	0.682	0.000	3	1.000	0.000	0.691	0.000
	90754789	7	0.999	0.001	0.670	0.001	3	1.000	0.000	0.690	0.000
	8988812	8	1.000	0.000	0.674	0.000	3	1.000	0.000	0.690	0.000
	2	4	1.000	0.000	0.683	0.000	4	1.000	0.000	0.690	0.000

Table A.9: All results from our adversarial setup on the synthetic datasets.

Dataset	Seed	LSTM					MLP				
		Epoch	F1	TVD	JSD	ATTEFA	Epoch	F1	TVD	JSD	ATTEFA
Diabetes	10	21	0.732	0.017	0.693	0.035	58	0.180	0.145	0.691	0.291
	50	25	0.730	0.018	0.693	0.037	76	0.203	0.142	0.691	0.286
	257	23	0.730	0.018	0.693	0.037	66	0.167	0.146	0.690	0.293
	500231	22	0.730	0.018	0.693	0.035	21	0.112	0.147	0.690	0.295
	100078	25	0.726	0.020	0.693	0.040	34	0.227	0.143	0.691	0.287
	12504	20	0.735	0.019	0.693	0.038	80	0.159	0.146	0.691	0.292
	90754789	31	0.728	0.018	0.693	0.035	24	0.151	0.145	0.691	0.291
	8988812	10	0.723	0.020	0.693	0.040	15	0.006	0.153	0.690	0.307
	2	18	0.729	0.017	0.692	0.034	5	0.000	0.153	0.691	0.308
Anemia	10	32	0.900	0.057	0.693	0.115	30	0.833	0.094	0.692	0.189
	50	14	0.877	0.072	0.693	0.144	13	0.842	0.086	0.692	0.173
	257	15	0.923	0.047	0.692	0.093	45	0.841	0.095	0.693	0.190
	500231	19	0.914	0.049	0.693	0.098	11	0.829	0.091	0.693	0.182
	100078	26	0.894	0.061	0.693	0.123	23	0.828	0.093	0.692	0.185
	12504	14	0.877	0.077	0.693	0.155	29	0.817	0.103	0.693	0.206
	90754789	26	0.912	0.052	0.693	0.104	19	0.835	0.093	0.692	0.186
	8988812	19	0.888	0.062	0.693	0.124	23	0.831	0.091	0.692	0.183
	2	17	0.927	0.046	0.692	0.092	14	0.834	0.091	0.693	0.182
SST	10	23	0.825	0.038	0.643	0.082	23	0.657	0.172	0.658	0.362
	50	16	0.822	0.033	0.624	0.072	10	0.598	0.172	0.656	0.363
	257	29	0.821	0.032	0.624	0.071	28	0.577	0.172	0.655	0.364
	500231	11	0.823	0.033	0.624	0.073	22	0.580	0.173	0.657	0.365
	100078	30	0.822	0.035	0.624	0.078	15	0.576	0.174	0.653	0.370
	12504	20	0.828	0.033	0.624	0.074	61	0.590	0.173	0.660	0.365
	90754789	18	0.818	0.035	0.624	0.078	21	0.630	0.173	0.657	0.364
	8988812	24	0.823	0.035	0.624	0.077	9	0.636	0.173	0.654	0.366
	2	18	0.823	0.035	0.624	0.078	14	0.599	0.172	0.656	0.363
IMDB	10	62	0.896	0.034	0.691	0.069	13	0.210	0.191	0.689	0.385
	50	42	0.889	0.040	0.689	0.081	18	0.230	0.188	0.689	0.379
	257	50	0.891	0.037	0.691	0.074	10	0.101	0.191	0.689	0.384
	500231	26	0.893	0.038	0.691	0.077	45	0.073	0.190	0.689	0.382
	100078	44	0.889	0.033	0.691	0.066	11	0.180	0.191	0.689	0.384
	12504	41	0.892	0.035	0.691	0.070	12	0.195	0.190	0.689	0.382
	90754789	60	0.890	0.040	0.691	0.081	29	0.100	0.191	0.690	0.384
	8988812	65	0.875	0.049	0.691	0.098	45	0.215	0.189	0.689	0.380
	2	54	0.890	0.038	0.691	0.077	10	0.120	0.191	0.689	0.385
AgNews	10	27	0.959	0.006	0.680	0.013	11	0.630	0.164	0.670	0.340
	50	28	0.958	0.006	0.681	0.013	17	0.567	0.174	0.671	0.359
	257	33	0.957	0.008	0.685	0.016	10	0.653	0.169	0.671	0.349
	500231	60	0.958	0.006	0.680	0.012	49	0.639	0.172	0.670	0.356
	100078	37	0.958	0.006	0.680	0.013	17	0.610	0.167	0.671	0.345
	12504	78	0.957	0.008	0.685	0.016	19	0.597	0.176	0.672	0.362
	90754789	68	0.959	0.007	0.685	0.015	27	0.633	0.171	0.671	0.352
	8988812	66	0.958	0.007	0.686	0.014	23	0.549	0.179	0.673	0.369
	2	47	0.958	0.006	0.680	0.012	39	0.611	0.179	0.671	0.370
20News	10	23	0.849	0.057	0.689	0.115	34	0.246	0.207	0.661	0.433
	50	26	0.863	0.038	0.690	0.077	34	0.176	0.200	0.653	0.425
	257	21	0.850	0.050	0.688	0.100	23	0.315	0.207	0.656	0.437
	500231	17	0.858	0.050	0.687	0.101	57	0.397	0.206	0.659	0.434
	100078	16	0.894	0.031	0.689	0.063	3	0.460	0.211	0.639	0.457
	12504	13	0.859	0.052	0.689	0.105	4	0.440	0.211	0.641	0.456
	90754789	16	0.874	0.049	0.688	0.099	44	0.043	0.205	0.654	0.435
	8988812	7	0.863	0.047	0.688	0.094	11	0.495	0.214	0.644	0.461
	2	19	0.875	0.042	0.689	0.084	10	0.492	0.213	0.642	0.461

Table A.10: All results from our adversarial setup on the real-life datasets.

Appendix B

Extended Early-Prediction Evaluation Results

This appendix complements Chapter 5 with detailed metrics (ERDE, latency-weighted F1, ranking-based evaluation). Tables B.1 and B.2 provide the metrics for the decision-based and the ranking-based metrics, respectively, for the early evaluation results on each task. The provided results follow a similar format as the overviews for the eRisk shared tasks (Losada et al., 2019, 2020; Parapar et al., 2021, 2022) to help assess our models compared to the shared task participants.

For details on the ERDE and the latency-weighted F1 (lw_{F1}) metric, refer to Section 3.4.1. The methodologies for the other evaluation metrics are described in the eRisk overview papers from 2019 to 2024 (Losada et al., 2019, 2020; Parapar et al., 2021, 2022, 2023, 2024).

Dataset	A	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	l_{TP}	$speed$	lw_{F1}
Anorexia	0.962	0.762	0.836	0.797	0.068	0.033	11.0	0.961	0.766
Depression	0.903	0.417	0.980	0.585	0.057	0.027	13.0	0.953	0.558
Gambling	0.984	0.746	1.000	0.855	0.021	0.008	3.0	0.992	0.848
Self-harm	0.903	0.522	0.842	0.645	0.084	0.046	14.0	0.949	0.612

Table B.1: Decision-based metrics of the early evaluation of our model on each task.

Dataset	1 post			10 posts			100 posts			500 posts			1000 posts		
	$P@10$	$NDCG@10$	$NDCG@100$	$P@10$	$NDCG@10$	$NDCG@100$	$P@10$	$NDCG@10$	$NDCG@100$	$P@10$	$NDCG@10$	$NDCG@100$	$P@10$	$NDCG@10$	$NDCG@100$
Anorexia	0.90	0.94	0.63	1.00	1.00	0.77	1.00	1.00	0.91	1.00	1.00	0.92	1.00	1.00	0.92
Depression	0.80	0.87	0.49	0.70	0.78	0.59	0.80	0.86	0.74	1.00	1.00	0.75	0.90	0.94	0.78
Gambling	1.00	1.00	0.67	1.00	1.00	0.85	1.00	1.00	0.95	1.00	1.00	0.99	1.00	1.00	0.99
Self-harm	1.00	1.00	0.66	0.90	0.92	0.72	1.00	1.00	0.88	1.00	1.00	0.91	1.00	1.00	0.92

Table B.2: Ranking-based metrics of the early evaluation of our model on each task.