

Erkennung von Depression und Angst auf Sozialen Medien unter Verwendung Selektiver Maskierung

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Data Science

eingereicht von

Princ Mullatahiri, B.Sc

Matrikelnummer 11846033

an der Fakultät für Informatik
der Technischen Universität Wien

Betreuung: Associate Prof. Dipl.-Ing. Dr.techn. Nysret Musliu

Mitwirkung: Dr. Hannah Metzler

B.Sc. M.Sc., Dr. Segun Taofeek Aroyehun

Wien, 7. Mai 2024



Princ Mullatahiri



Nysret Musliu

Detecting Depression and Anxiety on Social Media Using Selective Masking

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Data Science

by

Princ Mullatahiri, B.Sc

Registration Number 11846033

to the Faculty of Informatics

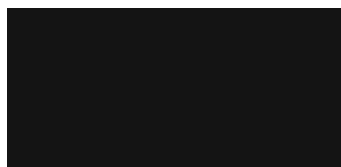
at the TU Wien

Advisor: Associate Prof. Dipl.-Ing. Dr.techn. Nysret Musliu

Assistance: Dr. Hannah Metzler

B.Sc. M.Sc., Dr. Segun Taofeek Aroyehun

Vienna, 7th May, 2024



Princ Mullatahiri



Nysret Musliu

Erklärung zur Verfassung der Arbeit

Princ Mullatahiri, B.Sc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 7. Mai 2024



Princ Mullatahiri

Danksagung

Mein aufrichtiger Dank gilt meinem Hauptbetreuer, Herrn Associate Prof. Dipl.-Ing. Dr.techn. Nysret Musliu, sowie meinen Co-Betreuern, Frau Dr. Hannah Metzler und Herrn B.Sc. M.Sc. Dr. Segun Taofeek Aroyehun. Ihre unermüdliche Unterstützung, ihr fachlicher Rat und ihre wertvollen Einblicke waren während meiner gesamten Forschungszeit von unschätzbarem Wert. Ihr Mentoring war entscheidend für die Gestaltung dieser Arbeit.

Mein aufrichtiger Dank gilt auch dem Team des Complexity Science Hub Vienna, insbesondere Prof. David Garcia und Dr. Max Pellert, für die Möglichkeit, mit ihnen zusammenzuarbeiten. Die Erkenntnisse, die ich aus dieser Zusammenarbeit gewonnen habe, sind von großer Bedeutung.

Mein besonderer Dank gilt meinem Mentor, Assoc. Prof. Dipl.-Ing. Dr.techn. Peter Knees, der mich in ethischen Fragen meiner Dissertation unterstützt und angeleitet hat. Seine fachliche Expertise und seine umsichtigen Ratschläge waren entscheidend für die Wahrung der ethischen Integrität meiner Forschung.

Schließlich möchte ich meiner Familie und meinen Freunden meinen tiefsten Dank aussprechen. Ihre fortwährende Unterstützung, Ermutigung und ihr Verständnis waren entscheidend für meinen Weg. Ihre Einsichten und ihr Zuspruch waren immer eine große Stütze, und ich bin zutiefst dankbar für ihr Engagement.

Acknowledgements

I would like to express my sincere gratitude to my main supervisor, Associate Prof. Dipl.-Ing. Dr.techn. Nysret Musliu and my co-supervisors, Dr. Hannah Metzler and B.Sc. M.Sc., Dr. Segun Taofeek Aroyehun, for their unwavering support, guidance and insights throughout the course of my research. Their mentorship has been instrumental in shaping this thesis.

I am deeply thankful to the Complexity Science Hub Vienna team, Prof. David Garcia and Dr. Max Pellert, for the invaluable opportunity to collaborate with them; I am grateful for the insights gained from our collaboration.

A special acknowledgment goes to my mentor Associate Prof. Dipl.-Ing. Dr.techn. Peter Knees for providing guidance and support in navigating ethical concerns related to my thesis. His expertise and thoughtful advice have been crucial in ensuring the ethical integrity of my research.

Last but not least, I want to thank my family and friends for their unwavering support, encouragement, and understanding. Their insights and encouragement have been crucial, and I am truly grateful for their involvement.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Kurzfassung

Psychische Gesundheitsprobleme sind eine der größten globalen Herausforderungen. Es wird geschätzt, dass mindestens jeder vierte Mensch einmal in seinem Leben von einer psychischen Störung betroffen ist. Depressionen sind dabei die häufigste Erkrankung, an der 5% der Erwachsenen im Laufe ihres Lebens leiden. Die Nutzung sozialer Medien hat in den letzten zehn Jahren stark zugenommen und stellt daher eine vielversprechende Datenquelle für die Schätzung psychischer Erkrankungen auf Bevölkerungsebene dar.

Unser Ziel war es, ein Modell zu entwickeln, das während des Domain-Specific Pre-Training (DSPT) eine selektive Maskierung anwendet, indem Wörter, die direkt mit Depression und Angst assoziiert sind, mit einer höheren Wahrscheinlichkeit maskiert werden. Dieser Ansatz ermöglichte es dem Modell, die charakteristischen Muster dieser Zustände besser zu erkennen.

Nach der Extraktion der Daten aus den verschiedenen Subreddits wurden die Daten mit Hilfe von Annotationen in Angst-, Depressions- und Zufallsklassen eingeteilt. Die Zufallsklasse umfasste Daten aus 16 verschiedenen Subreddits, einschließlich der beliebtesten. Wir sammelten Twitter-Daten, indem wir nach Tweets suchten, die öffentlich über Depression oder Angst berichteten, und ordneten sie den Klassen Angst oder Depression zu. Die Daten für die Zufallsklasse wurden aus den Archiven des Complexity Science Hub Vienna gewonnen. Für die Daten aus Reddit und Twitter wurden verschiedene Vorverarbeitungsschritte zur Qualitätssicherung durchgeführt.

Wir haben verschiedene Strategien implementiert, um die wichtigsten Wörter und ihre zugehörigen Maskierungswahrscheinlichkeiten für die selektive Maskierung während der DSPT zu identifizieren. Diese Strategien umfassten überwachtes Lernen, Clustering, Log-Odds, Term Frequency-Inverse Document Frequency (TF-IDF) und manuell ausgewählte Wörter. Die Modelle wurden mit Daten aus Reddit trainiert. In der Feinabstimmungsphase haben wir Daten von Reddit für das Training und Daten von Twitter für die Evaluierung verwendet, was uns geholfen hat, ein Modell zu entwickeln, das gut auf Daten von verschiedenen Social Media Plattformen verallgemeinert werden kann.

Wir kamen zu dem Schluss, dass die selektive Maskierung von Wörtern, die direkt mit Depression oder Angst in Verbindung gebracht werden, besonders effektiv ist, um das Auftreten falsch negativer Ergebnisse zu minimieren. Angesichts unseres Ziels, ein Modell mit einer minimalen Anzahl von falsch-positiven und einer geringen Anzahl von

falsch-negativen Ergebnissen zu entwickeln, verwendeten wir den F1-Score zur Bewertung. XGBoost- und Clustering-Strategien erwiesen sich als die leistungsfähigsten Strategien für die selektive Maskierung und zeigten nicht nur gute Ergebnisse, sondern auch Stabilität. Der höchste erreichte F1-Score betrug 0,8137 für Depression und 0,9236 für Angst und übertraf damit die Basismodelle mit Werten von 0,7504 für Depression und 0,8965 für Angst.

Angesichts des Black-Box-Charakters und der eingeschränkten Interpretierbarkeit aktueller Modelle haben wir die Transparenz und Interpretierbarkeit durch die Integration globaler und lokaler Erklärungstechniken wie Local Interpretable Model-agnostic Explanations (LIME) und Shapley Additive Explanations (SHAP) verbessert. Unsere Ergebnisse deuten darauf hin, dass Wörter wie Pronomen der ersten Person, Schimpfwörter und Wörter, die mit dem Ausdruck von Emotionen assoziiert werden, signifikant zu positiven Vorhersagen beitragen. Zusammenfassend lässt sich sagen, dass wir ein Modell entwickelt haben, das sich gut auf Daten von verschiedenen Social Media Plattformen verallgemeinern lässt, verbesserte Ergebnisse liefert und mehr Transparenz bietet.

Abstract

Mental health problems are one of the major problems in the world. It is estimated that once in their life, at least one mental health condition will affect one in four people. Depression is the most common condition, with 5% of adults suffering from it in their lifetime [1]. The use of social media has grown significantly in the last decade, making social media a promising source of data to estimate mental health conditions at the population level.

We aimed at developing a model that employed selective masking during Domain-Specific Pre-Training (DSPT), where words directly linked to depression and anxiety were given a higher probability of getting masked. This approach enabled the model to better understand the distinctive pattern characteristics of these conditions.

After extracting the data from various subreddits on Reddit, we applied annotations to classify the data into anxiety, depression, or random classes. The random class incorporated data from 16 subreddits, including the most popular ones. We collected Twitter data by searching for tweets featuring public self-disclosure of depression or anxiety diagnoses. These tweets were annotated under the anxiety or depression class. Random class data were compiled from Complexity Science Hub Vienna archives. We implemented various pre-processing steps for Reddit and Twitter to ensure data quality.

We employed various strategies to identify the most meaningful words and their associated masking probabilities for selective masking, used during DSPT. These strategies were supervised learning models, clustering, log-odds, Term Frequency-Inverse Document Frequency (TF-IDF), and manually selected words. We trained the models using data from Reddit. Furthermore, during the fine-tuning phase, we used Reddit data for training and Twitter data for evaluating, contributing to developing a model which generalized well on data from different social media platforms.

We concluded that selective masking of words directly linked with depression or anxiety proved particularly effective in minimizing the occurrence of false negatives. Given our interest in developing a model with minimal false positives, as well as a small number of false negatives, we used the f1-score for evaluation. XGBoost and clustering strategies emerged as the best-performing strategies for selective masking, demonstrating not only good results but also stability. The highest achieved f1-score for the depression domain was 0.8137, and for the anxiety domain, 0.9236, surpassing baseline model scores of 0.7504 for the depression domain and 0.8965 for the anxiety domain.

Given state-of-the-art models' black-box nature and limited interpretability, we enhanced transparency and interpretability by incorporating global and local explainability techniques such as Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP). Our findings suggested that words like first-person pronouns, curse words, and words related to expressing emotions significantly contributed to positive predictions. In summary, we developed a model that generalized well on data from different social media platforms, produced improved results, and had a higher transparency.

Contents

Kurzfassung	xi
Abstract	xiii
Contents	xv
1 Introduction	1
1.1 Problem statement	1
1.2 Thesis Goals and Contributions	2
1.3 Methodology	3
1.4 Organization	4
2 Theoretical Foundations of Domain-Specific Pre-Training	5
2.1 Development of Natural Language Processing	5
2.2 BERT and RoBERTa	8
2.3 Domain-Specific Pre-Training	10
3 Proposed Architecture and Implementation	13
3.1 Data Collection and Annotation	13
3.2 Ethics and Limitations	17
3.3 Data Pre-Processing	19
3.4 Meaningful words detection	21
3.5 Domain Specific Pre-Training	22
3.6 Fine-Tuning	25
4 Results	27
4.1 Depression Domain	28
4.2 Anxiety Domain	34
4.3 Comparative Analysis and Explainability	38
5 Conclusion	47
5.1 Main Conclusions	47
5.2 Contribution to the state-of-the-art	49
	xv

A Appendix	52
A.1 Figures	52
A.2 Tables	64
List of Figures	69
List of Tables	71
Bibliography	75

Introduction

Mental health problems are one of the major problems in the world. It is estimated that at least one mental health condition will affect one in four people once in their lifetime. According to statistics by World Health Organization (WHO), depression and anxiety disorders are among the most prevalent mental health conditions, where approximately 4.4% of the population is currently suffering from depression, while 3.6% are affected by anxiety disorders. Both disorders are more common for females than males [1]. The COVID-19 pandemic triggered an increase of 27.6% in the prevalence of depression and a 25.6% increase in the prevalence of anxiety worldwide [2], making mental health problems a significant and pressing public health concern.

1.1 Problem statement

Rates of depression, anxiety, and suicidal thoughts in the population are influenced by societal events and trends such as pandemics, economic crises including unemployment, or missing positive perspectives for the future, such as those around climate change [3, 4, 5]. Addressing these problems via public health interventions requires an understanding of how specific mental health issues are related to such events, as well as mental health estimates at the population level. Traditionally, such estimates are collected with representative surveys of the population. Yet, such surveys require a lot of resources, and it is expensive to conduct them regularly and impossible to conduct them in real-time.

Recently, traces of behaviour on social media have become a promising source of data to estimate mental health conditions at the population level. Machine Learning (ML) models and Natural Language Processing (NLP) are promising tools that allow detecting indicators of different mental health issues at the macro-scale [6]. However, there are currently very few publicly available models for the detection of depression and anxiety. Standard NLP models are trained on general online text data, whereas mental health

problems are a very specific domain, making the performance of available models in this field inferior. Better models are required to improve research on the relationship between mental health and societal trends.

1.2 Thesis Goals and Contributions

The latest advances in pre-trained contextualized language representations have given rise to the development of several domain-specific pre-trained models [7, 8, 9, 10]. Considering the performance of domain-specific pre-training in other domains, we aimed to use this approach to improve the state-of-the-art models on the detection of depression and anxiety from public social media posts, thus releasing two models trained on these specific domains for the detection of depression and anxiety at the macro-scale.

To do so, we used selective masking, where words related to the mental health disorder in question had a higher probability of getting masked compared to the other words. Thereby, the model learned more about the patterns associated with the specific condition. Figure 1.1 presents the proposed architecture for adding a DSPT using data gathered from Reddit between general pre-training (based on data from Wikipedia and Bookcorpus) and fine-tuning.

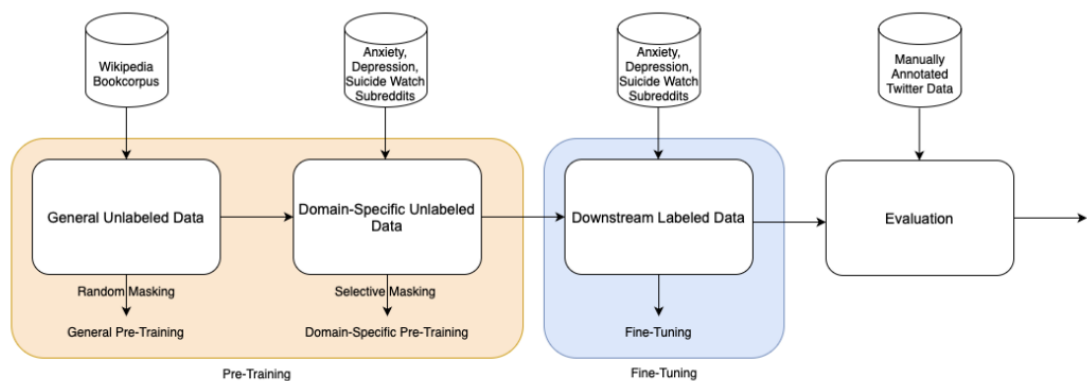


Figure 1.1: Proposed Architecture: Add DSPT between general pre-training and fine-tuning

The main research questions of this theses are:

1. How does pre-training with domain-specific unlabeled data influence the model's performance?
2. How does selective masking of words specifically linked to depression or anxiety improve the model's performance?
3. How well does the model generalize when using Reddit data for training and Twitter data for evaluation?

The main contributions of this thesis were:

- A model that generalized well on data from different social media platforms: Through cross-platform evaluation, this thesis aimed to develop a model that can be applied to other social media platforms without a significant loss in accuracy.
- Improved accuracy: The development of a domain-specific pre-trained model that captured mental health disorder patterns achieved better results than current state-of-the-art models.
- Improved global and local explainability: By using both global and local explanation methods, this thesis aimed to provide insight into the reasoning behind the model's predictions, enhancing its interpretability and transparency.

These contributions aimed to address some of the existing limitations and challenges in the detection of mental health disorders at the population scale.

1.3 Methodology

The proposed solution was to add a DSPT between general pre-training and fine-tuning. The methodology included the following stages:

Data gathering: For the general pre-training phase, we used data also used by other algorithms like Bidirectional Encoder Representations from Transformers (BERT)[11] and Robustly Optimized Bidirectional Encoder Representations from Transformers (RoBERTa)[12], including English Wikipedia and Bookcorpus. For DSPT, we used data from different subreddits, grouping them into control and treatment datasets. The control dataset consisted of data from the most popular subreddits, whereas the treatment dataset contained data from the anxiety and depression subreddits, respectively. We gathered data from Twitter using the same method as the CLPsych Shared Task 2015 dataset [13], where we searched tweets that contained a public self-disclosure of a depression or anxiety diagnosis. Then researchers at Complexity Science Hub Vienna (CSH) manually checked these tweets to remove sarcastic and other non-related tweets. Afterwards, we removed the keywords used for searching these tweets to avoid biasing model performance.

Pre-processing: We removed deleted posts, handled slang words, and removed hashtags and user mentions. We applied pre-processing steps to both the Reddit and Twitter datasets.

Meaningful words detection: for this step, we used two methods:

- We trained a supervised learning estimator with a fit method that provided information about feature importance, such as linearSVC [14] and XGBoost [15]. We then obtained the most important features from that model, specifically the most meaningful words for that specific domain.

- We chose meaningful words based on the expert knowledge of a psychological researcher (Hannah Metzler) with a track record in mental health research.

DSPT: We used selective masking during DSPT, where more meaningful words had a higher chance of getting masked compared to other words. To overcome the limitation of the small number of words that the Huggingface build-in tokenizer [16] was trained on, we used whole word masking [17].

Fine-Tuning: We used data gathered from Reddit to fine-tune the model for the main task. Compared to the pre-training phase, this phase was inexpensive regarding time and resources.

Evaluation: We evaluated the models on data gathered from Twitter, in order to assess how well the model generalizes to data from a different social media platform.

Interpretability and Transparency: Although Deep Learning (DL) NLP models achieve some of the highest performances to predict mental health disorders from text, they had the disadvantage of being black-box models where the reasons for a specific label were not transparent. Increasing the explainability of such models was important to check if predictions were based on plausible patterns and to understand better what types of patterns in data indicated mental health issues. Hence, to make the model more explainable, we used LIME [18] and SHAP [19]. In this thesis, we also address ethical considerations within the Proposed Architecture and Implementation section.

1.4 Organization

The second chapter discusses the theoretical foundations of DSPT drawing insights from existing literature and comparable studies. The third chapter discusses the proposed architecture including data collection and ethical considerations, data pre-processing, meaningful words detection, DSPT and fine-tuning. The fourth chapter presents the outcomes of employing selective masking strategies using both BERT and RoBERTa as base models for predictions in the depression and anxiety datasets. Finally in fifth chapter we explore key findings and elaborate on our contributions to the current state-of-the-art models.

Theoretical Foundations of Domain-Specific Pre-Training

The theoretical foundations of DSPT are presented in this chapter, which also explores other studies focused on creating ML models in the mental health domain.

2.1 Development of Natural Language Processing

NLP is a subset of Artificial Intelligence (AI) that focuses on the interaction between computers and languages. NLP is focused on development of algorithms that help computers to interact with human language in a way that is understandable to humans. In recent years, the field of NLP has been rapidly evolving. With the recent advances in DL and large-scale language models, computers have achieved human-like performances and sometimes even outperformed humans on many NLP tasks. NLP has a wide range of applications starting from language translation, sentiment analysis, chatbots and speech recognition.

Even before the development of state-of-the-art NLP methods, such as BERT and RoBERTa, a significant amount of research in NLP was focused on the mental health domain. Researchers aimed at developing models that could aid in the detection of mental health issues, for example, to identify which medications were proven to be most effective [20].

Given that most models only operate with numerical data, translating human language to machine language is an essential step. Term frequency was used for this purpose. However, term frequency had limitations, as frequently used words often lacked significant representative meaning. To overcome this limitation, researchers have used alternative methods such as:

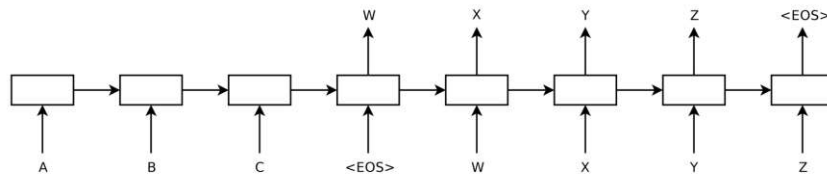
- TF-IDF: gives a higher weight to the important words in a document by taking into account their frequency in the document and rarity in the entire corpus. It is one of the most prevalent metrics used in text-based recommender systems. Where term-frequency is calculated using the following formula: $tf(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$, where $f_{t,d}$ is the number of times that word t occurs in document d , and the denominator is the total number of words in document d . Whereas inverse-document-frequency is calculated using the following formula: $idf(t, D) = \log \frac{N}{n_t}$ where N is the total number of documents and n_t is the number of documents where term t appears. The formula for TF-IDF is the following:

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D)$$

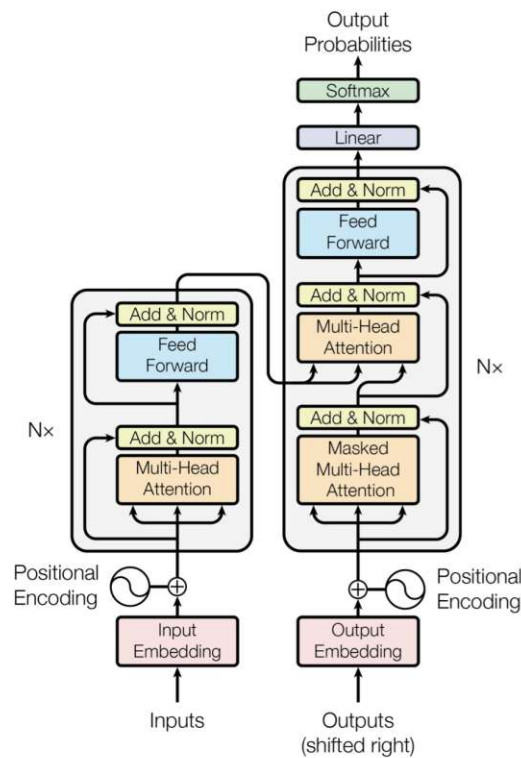
- Word2Vec: is a method for learning vector representations of words. Word2Vec captures the meaning of words based on their usage on large texts, it works by training a neural network to predict neighbouring words of a given word within a text.
- Best Matching 25 (BM25): it is based on TF-IDF of the terms in the document, where it also considers the document length and average document length in the corpus.

Support Vector Machine (SVM) have been widely used by researchers as a model of choice for NLP tasks after performing data pre-processing. Given the simplicity of SVMs, they have produced satisfactory results. With the growth of DL, more complex models have been used in the mental health domain. These models ranged from simple neural network models to Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) models. One of the challenges of RNNs was vanishing and exploding gradients. In a vanishing gradient, the gradient signal weakened as it propagated through the layers, making it challenging to learn long-term dependencies. Conversely, in exploding gradients, the network's weight updates were so large that the model overshoot the optimal values, resulting in difficulties converging to a good solution. LSTM models addressed this issue by utilizing a gating mechanism composed of three gates: input, forget and output gate. This gating mechanism enabled the neural network to selectively retain or discard information from previous steps, effectively handling long-term dependencies. One of the widely used approaches for NLP tasks was Seq2Seq Learning with LSTM [21]. This approach used two LSTMs, an encoder network and a decoder network. The encoder read the input sequence, one time-step at a time, and it obtained a large vector representation of fixed dimensions, whereas the decoder extracted the output sequence from that vector representation. The sequential dependency between time-steps can be seen in Figure 2.1(a) where given an input sentence "ABC", it reads it one time-step at a time, and then the output of each time-step is fed back into the decoder, which generates the next word until it outputs the end-of-sentence token. The limitations of using LSTMs were the difficulty in handling long input sequences as well as the sequential dependency

between time-steps where the output at each time-step was conditioned on the previous output.



((a)) Seq2Seq LSTM structure



((b)) Transformer model architecture

Figure 2.1: Seq2Seq structure [21] and Transformers architecture [22]

To overcome these limitations, transformers [22] were introduced. They used a self-attention mechanism that allowed the model to attend to different parts of the input sequence at different time-steps. The self-attention mechanism allowed the model to handle long input sequences while processing them in parallel, making it more efficient than Seq2Seq models. Figure 2.1(b) presents the architecture of transformers. Positional encoding was used to capture information about the position of each word in the input sentence. Additionally, they used an encoder and a decoder, each consisting of a stack of identical layers. Each layer had two sub-layers, a self-attention mechanism and a

position-wise feed-forward neural network. Multi-head attention was a critical component within transformers. The attention function was defined as mapping a query and a set of key-value pairs to an output, as depicted in Figure 2.2. Multiple attention heads were created, and each had its own set of learned weights; this enabled the model to focus on different relationships in the input sequence and made it possible to understand complex relationships in the input data. The only difference with masked multi-head attention is that it used a masking mechanism that prevented the model from attending to future positions in the input sequence during training. Overall, transformers laid the foundation for a new era of attention-based neural network architectures such as BERT and RoBERTa.

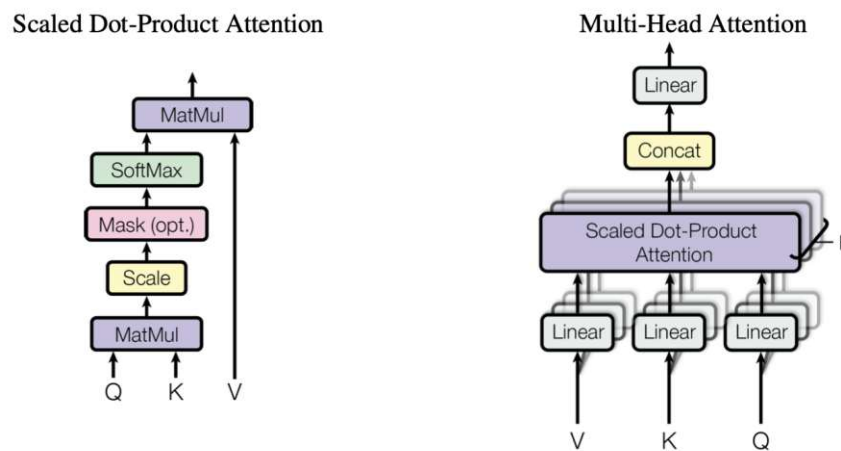


Figure 2.2: Scaled Dot-Product Attention and Multi-Head Attention of transformers [22]

2.2 BERT and RoBERTa

BERT [11] was designed to pre-train deep bidirectional representations from the unlabeled text by simultaneously conditioning on both left and right context across all layers. The pre-trained BERT model could be fine-tuned with just an additional output layer to develop state-of-the-art models for various tasks such as chatbots and text classification without the need for task specific modifications in the model’s architecture. BERT used a multi-layer bidirectional transformer encoder, which consisted of two steps in its architecture: pre-training and fine-tuning, as can be seen in Figure 2.3. During the pre-training phase, the model was trained on unlabeled data such as English Wikipedia and Bookcorpus [23] over different pre-training tasks. During the fine-tuning phase, BERT model underwent initialization with pre-trained parameters, followed by the fine-tuning of all these parameters using the labelled data from the specific task.

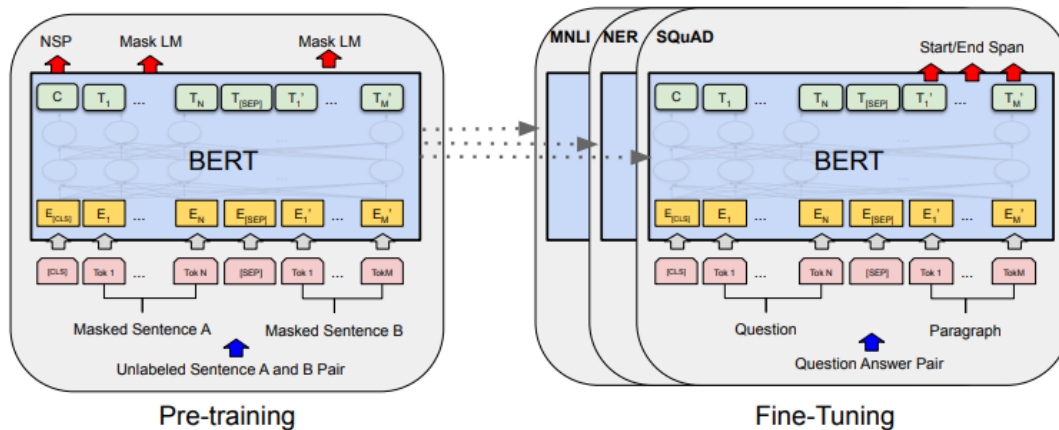


Figure 2.3: BERT architecture: pre-training and fine-tuning [11]

In their paper on BERT, Devlin et al. introduced two models: BERT Base, comprising 12 transformer blocks with a hidden size of 768 and 12 self-attention heads, totaling 110M parameters, and BERT Large, featuring 24 transformer blocks with a hidden size of 1024, 16 self-attention heads, and 340M parameters. BERT was trained on two unsupervised tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). To train a deep bidirectional representation, a masking strategy was employed in BERT. Specifically, 15% of input tokens were randomly masked, and then the model aimed to predict these masked tokens. To mitigate the [MASK] token not appearing during the fine-tuning phase, out of the 15% input tokens that were selected to be masked, 80% of the time that token was replaced with [MASK], 10% it was replaced with a random token, and 10% was left unchanged. For BERT to understand the sentence relationships, the model was pre-trained for a binary NSP task. Specifically, when choosing the sentences A and B for each pre-training example, 50% of the time, B was the following sentence, and the other 50% B was a random sentence from the corpus.

RoBERTa [12] was a replication study of BERT that examined the influence of several key hyperparameters and the size of training data. The modifications implemented in RoBERTa included: training the model for extended durations with more data and with larger batches, removing the NSP objective, training on longer sequences, and dynamically changing the masking pattern for each epoch which was applied to the training dataset.

RoBERTa was trained on multiple text corpora: Bookcorpus, English Wikipedia, Cc-News [24], OpenWebText [25] and Stories [26]. BERT implementation performed masking once during data pre-processing, which resulted in a single static mask, whereas RoBERTa implementation, instead of using a fixed masking for the entire training duration, generated a fresh masking pattern for each batch of the training example. RoBERTa used byte-level Byte Pair Encoding (BPE) vocabulary of 50K size compared to word piece tokenization of 30k size that BERT used. BPE further tokenized words into sub-word units, which

helped the model handle out-of-vocabulary words. Byte-level BPE was particularly useful for handling multilingual text, as different languages have different encoding schemes. Overall, these models have revolutionised language understanding, paving the way for significant advancements in NLP tasks.

2.3 Domain-Specific Pre-Training

BERT performed pre-training using general data from English Wikipedia and Bookcorpus. However, the pre-training did not consider the downstream task to which the model would be applied. This limitation emphasized the significance of DSPT as an additional step in the model's training pipeline. DSPT referred to the process of training the model on domain-specific data to enhance its understanding of specific patterns in that particular domain. In recent years, there have been published a considerable amount of papers focused on DSPT. In Gururangan et al. [27], they showed that learning domain-specific and task-specific patterns during pre-training led to performance gains even in scenarios with limited resources. They concluded that adapting to the task-specific unlabeled data improved performance even after DSPT. The difference between DSPT and Task-Specific Pre-Training (TSPT) was the size of unlabeled data, where TSPT had a smaller pre-training corpus but was much more task-relevant. In Gu et al. [28] they proposed a new approach which involved using selective masking in the added stage of task-guided pre-training which was between general pre-training and fine-tuning. Where in the task-guided pre-training stage, the model was trained using MLM on medium sized domain-specific unlabelled data, which consisted of other corpora in the same domain. During this stage, a selective masking strategy was used to focus on masking the important tokens. For the selective masking strategy, they defined a task-specific score for each token, where if the score was lower than a threshold, they regarded the token as important. In Sosea et al. [7], they applied DSPT to the model to help it learn emotion-related tasks. Emotion-related words had a probability of 50% of getting masked compared to other words, which had a probability lower than 15% of getting masked. They managed to improve the downstream performance with an average f1-score increase of 1.2%.

The main limitation of DSPT relied on finding the most important tokens to use for the selective masking task. To overcome this limitation, Arefyev et al. [29] proposed a technique for a more efficient way in finding these important tokens which relied on words with a higher weight of the Naive Bayes classifier trained for the specific task, these words were more relevant compared to most frequent words that other MLM models used. Through their research, the authors have demonstrated that their proposed technique provided faster adaptation and better performance for sentiment analysis. In the other relevant paper, Ji et al. [30] they used DSPT to aid the language model in understanding the patterns of Mental Health domain. They released two models, MentalBERT and MentalRoBERTa, which involved an additional pre-training stage specifically designed for BERT and RoBERTa architectures. The author's findings indicated that continued

pre-training with mental health-related corpus improved classification performance for mental health disorders.

This thesis examined the implementation of DSPT to enhance results within the domain of mental health disorders. MentalBERT and MentalRoBERTa focused on pre-training the language model using domain data from multiple mental health problems. However, recognizing that each mental health problem possesses its own unique patterns, we contended that better results could be achieved by employing selective masking techniques and using data that are specifically associated with individual mental health issues.

Proposed Architecture and Implementation

We present the proposed architecture and implementation in this chapter. Here, we outline the technical aspects of our solution, including the software tools, programming languages, and technologies employed. We intend to add another stage to BERT and RoBERTa between general pre-training and fine-tuning, which is presented in Figure 1.1.

In pre-training BERT and RoBERTa, general unlabeled data from English Wikipedia and Bookcorpus were used for training the model. In DSPT, we only used data specifically linked to depression and anxiety collected from Reddit. In fine-tuning, we used the downstream labelled data for binary classification. We evaluated the model's results on data collected from Twitter.

3.1 Data Collection and Annotation

In this section, we focus on the critical aspects of data collection and annotation, which are fundamental parts of our research. We collected data from two social media platforms: Reddit and Twitter. Reddit is a social network, where registered users submit posts such as texts, images, and links into user-created communities called subreddits. In the majority of subreddits, the primary language is English. This makes Reddit helpful in generating datasets for performing natural language processing tasks for English texts. Reddit contains multiple subreddits for a lot of communities, including communities for people who are diagnosed with depression and people who are diagnosed with anxiety. Considering that the number of collected data was too large for manual annotation, we used distant supervision, annotating data based on the respective subreddit. This approach is suitable since each subreddit has rules and moderators who delete posts from users which are not related to the subreddit's theme or break any rules. For both

3. PROPOSED ARCHITECTURE AND IMPLEMENTATION

mental health disorders, we created two datasets to train the models: the treatment and the control dataset. We collected the data from Reddit using RedditAPI [31] and Python. The treatment dataset comprised data sourced from the depression or anxiety subreddits, respectively. The control dataset consisted of data gathered from the most popular subreddits such as r/AskReddit, r/Aww, r/Books, r/ChangeMyView, r/Europe, r/Funny, r/GetMotivated, r/MadeMeSmile, r/Motivation, r/Movies, r/OutOfTheLoop, r/Politics, r/Technology, r/TodayILearned. The collected data, including the number of posts from each subreddit and the average amount of words per post, is presented in Table 3.1.

Subreddit	Nr. of Posts	Average Words per Post
Anxiety	162,956	177.07
AskReddit	5,353,434	17.80
Aww	520,326	16.90
Books	58,633	127.20
ChangeMyView	30,102	277.10
Depression	416,812	194.50
Europe	83,628	24.40
Funny	322,847	18.30
GetMotivated	30,118	49.00
InterestingAsF***	106,092	19.20
MadeMeSmile	56,992	30.70
Motivation	14,898	56.00
Movies	201472	61.30
NextF***ingLevel	42512	18.00
OutOfTheLoop	30331	53.10
Politics	519461	15.30
Technology	91092	20.99
TodayILearned	257924	29.60

Table 3.1: Number of posts and average number of words per post for each subreddit data

Twitter is a social media platform that allows users to share short messages, known as tweets. We collected data from Twitter using the same method as the CLPsych Shared Task 2015 dataset [13]. On Twitter, users often engage in public discourse regarding their health for different purposes, including seeking treatment or health-related guidance. Particularly for mental health, users often opt for a public platform such as Twitter as a means to challenge the negative association with mental illnesses. A significant number of Twitter users openly disclose their diagnosis, such as "I have been diagnosed with depression/anxiety ...". We searched these tweets that contained a public self-disclosure of a depression or anxiety diagnosis and then researchers at CSH manually checked them to remove sarcastic tweets and other non-related tweets. Afterwards, we removed words

used to search these tweets to avoid biasing model performance. Similar to the data collected from Reddit, we created two datasets using data gathered from Twitter: the treatment and control datasets. The treatment dataset contained tweets with public self-disclosure of depression or anxiety diagnosis, whereas the control dataset contained randomly collected tweets. Table 3.2 shows the number of tweets and the average words per tweet for each treatment/control dataset created from data collected from Twitter. To gain a deeper insight into the distinctions among different collected corpora, we conducted an analysis utilizing Shannon Entropy Shifts. In Figure 3.1 can be seen the graph from Shannon Entropy Shift between data collected from Reddit and data that BERT used in the general pre-training phase, and in Figure 3.2 can be seen the Shannon entropy shift between data collected from Reddit and data gathered from Twitter for depression dataset. We computed the same graphs for the anxiety datasets, which can be found in the appendix. We created all these graphs using shifterator library from Python, which was based on generalized word shift graphs [32]. Shannon Entropy Shift tries to find more surprising words and how they vary between two corpora, where the less often a word appears in a corpus, the more surprising that word is. Shannon Entropy H is calculated as:

$$H(P) = \sum_i p_i \log \frac{1}{p_i}$$

Where $\log \frac{1}{p_i}$ is the surprisal of a word, and p_i is the relative frequency of the word. We compared two texts by finding the difference between their entropy's:

$$\delta H_i = H(P^{(2)}) - H(P^{(1)})$$

If the result δH_i was positive, then word i had a higher score in the second corpus if it was negative then word i had a higher score in the first corpus. Figure 3.1 illustrates that words used in Wikipedia and Bookcorpus were less predictable than those used in Reddit. The lower-left quadrant of the word shift graph revealed that the top 50 words account for approximately 20% of the entropy difference between the two corpora. Furthermore, the analysis showed a notable prevalence of first-person pronouns and affective lexicon within the Reddit dataset. Additionally, an abundance of absolutist terms, such as "never", "anything", and "always", were observed more in the Reddit dataset. These words were more common on posts related to mental health problems [33]. Given these distinctive linguistic characteristics between these two corpora, we hypothesized that further pre-training using BERT and RoBERTa on data collected from Reddit could enhance their ability to perceive intricate linguistic patterns in posts related to mental health problems. Figure 3.2 reveals that the Reddit dataset contained more unpredictable words than the Twitter dataset, where the top 50 words accounted for approximately 17% of the difference in entropy between the corpora. The analysis of Reddit data indicated a greater prevalence of absolutist words, while the Twitter data exhibited a higher frequency of terms associated with medicinal references to mental health conditions. Furthermore, first-person pronouns did not feature among the top 50 words with the most significant entropy disparity, indicating that they were widespread across both corpora.

3. PROPOSED ARCHITECTURE AND IMPLEMENTATION

Tweets	Number of Posts	Average Words per Post
Anxiety	86,881	29.4
Depression	124,978	28.7
Random	763,704	17.1

Table 3.2: Number of tweets and average number of words per tweet

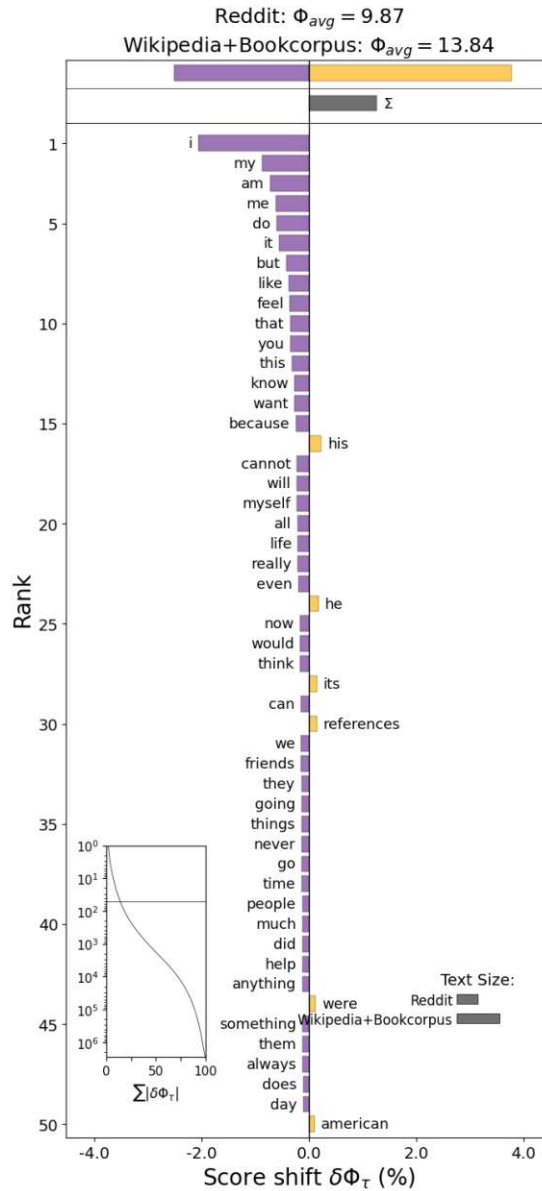


Figure 3.1: Shannon Entropy Shifts between Reddit data and Wikipedia+Bookcorpus

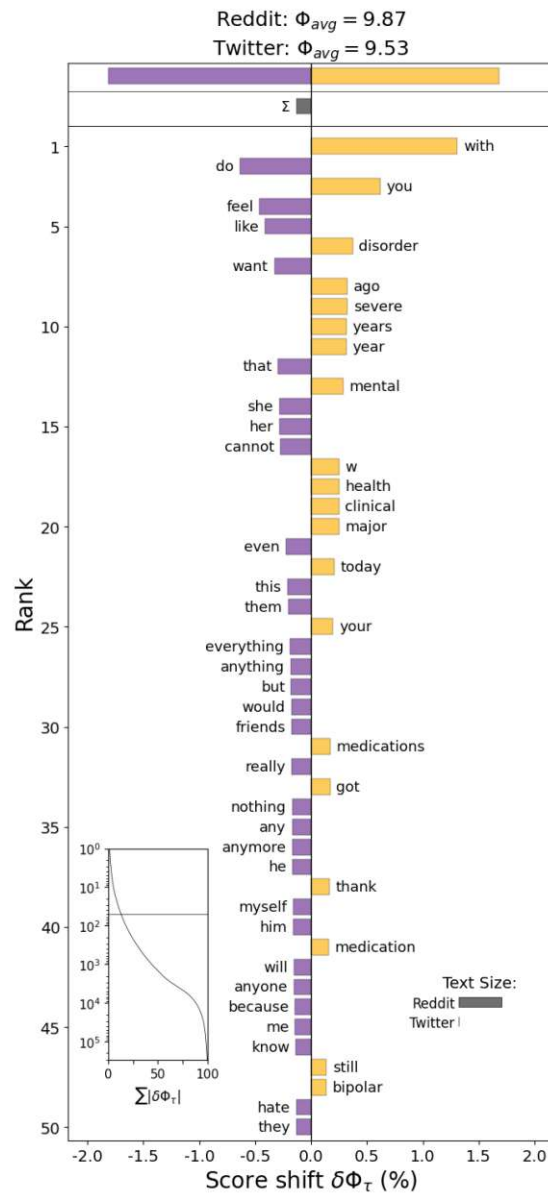


Figure 3.2: Shannon Entropy Shifts between Reddit and Twitter data

3.2 Ethics and Limitations

While this thesis achieved promising results in predicting depression and anxiety, it also considered several limitations. One of the limitations was the need for high computational power for pre-training the models. Pre-training these models required a Graphics Processing Unit (GPU) server and a large GPU memory, which limited the size of the dataset which could be used for the analysis. Additionally, this limitation affected the

exploration of additional model architectures and hyperparameters, which could have further improved the results. Another limitation is due to the high amount of data gathered from Reddit, which made manual annotation of each post infeasible; hence, there might have been some cases where training data was noisy, potentially limiting the accuracy of the models.

Furthermore, the usage of data from social media platforms raised ethical concerns. They regarded the data collection and sharing, the level at which we labelled the data for mental health disorders (individual or group-level), and the explainability of the models we trained.

Another limitation was the potential for adversaries to use the released model for malicious purposes. The Management Plan subsection addresses risk mitigation and ethical concerns and ensures the implementation of ethical standards measures.

Future work could address these limitations by exploring additional model architectures and hyperparameters and using better data annotation methods to enhance the model's accuracy and robustness.

3.2.1 Management Plan

We proposed a data and model management plan to address ethical issues. They assisted in building a methodology for handling the data and the models while minimising data ethics concerns like privacy violations, up-to-date data, and misuse of models. Using these protocols, we aimed to further improve the quality and reliability of results.

Data Management Plan

For this research, we gathered data from both Reddit and Twitter. We collected data from Reddit using Reddit API, while data from Twitter using tweepy library following the same method as the CLPsych Shared Task 2015 dataset. We searched for tweets containing public self-disclosure of depression or anxiety diagnosis, and then researchers at CSH manually checked these tweets to remove sarcastic tweets and other non-related tweets.

Developers using the Twitter API are required by the Twitter Developer policy to make user privacy a high priority. We refrained from gathering any personally identifiable data to uphold this policy. Chapter 3, subsection "Public display of Tweets and Twitter Content redistribution", specified that Twitter requires the use of the most current version of tweets when being displayed and the sharing of only IDs when the data are being redistributed. To ensure compliance with these policies, if needed, we will only share IDs of tweets [34]. Furthermore, Twitter's restricted use policy for the Twitter API permitted aggregate analysis of Twitter content provided that no personal data were stored [35].

Similarly, the Reddit Developer policy requires using the most current version of Reddit posts. However, developers were not permitted to use Reddit services/data for training ML models without prior permission. To comply with this policy, we submitted a formal

request, and Reddit admins subsequently approved it, as documented in the appendix of this thesis in Figure A.1 [36].

For both training and evaluation, we used only publicly available data, and we stored only IDs and text of Reddit/Twitter posts. We don't intend to release the data we obtained for this research since we value privacy and confidentiality. However, if publishing our results calls for data sharing to assure reproducibility, we will only share the IDs of these posts whilst following similar procedures with getting the CLPsych Shared Task 2015 dataset [37]. To guarantee ethical and privacy standards, we created a data use agreement with users seeking access to the data. This agreement specifies that users can only use the data for research purposes, that they can't share the data with other entities and that they will not attempt to violate any ethical and privacy standards. Only sharing the IDs of these posts ensures that when a user decides to delete their post from either Reddit or Twitter, those posts can't be retrieved via the ID anymore by other researchers. It also ensures that only posts that are still publicly available and their latest version can be accessed. Regarding mental health labels, we labelled Reddit posts only at the group level, not at an individual level, based on the subreddit that posts belonged to, whereas for tweets, labels were based on the public self-diagnosis of users on the tweet level.

Model Management Plan

Sharing of our models and their results was in line with ethical data sharing standards, as Article 89, Recital 162 of General Data Protection Regulation (GDPR) [38] specified that the result of processing for statistical purposes was not personal data but aggregate data. Therefore, we aimed to share the models generated during this research in a public repository on the HuggingFace platform. This allows other users to easily use the released models in their research and makes it possible for them to generate further improved models with higher accuracy and better interpretability. If needed, in case of malicious uses by adversaries, we can switch to a private repository or delete the model from the platform. By making these models publicly available, we aimed to contribute to the advancement of depression and anxiety detection research. To ensure transparency of the models, especially when dealing with black-box models, we employed both global and local explainable techniques. We aimed to learn more about the patterns in the data that might indicate a mental health problem and verify that believable patterns supported our predictions. To achieve this, we used methods such as LIME and SHAP. These techniques helped us to make the predictions of our models more transparent and interpretable.

3.3 Data Pre-Processing

Applying pre-processing steps was essential to ensure data consistency and mitigate potential biases embedded within hashtags or some specific words. As can be seen from Figure 3.3, the pre-processing procedures for data from Reddit encompassed eliminating posts that have been deleted by moderators or removed by users. Notably, since contextual

3. PROPOSED ARCHITECTURE AND IMPLEMENTATION

information can be concentrated either within the post title or the post body, we combined the title and body. To ensure data uniformity, we removed extraneous spaces, hashtags and hyperlinks which were regarded as devoid of meaningful information. Furthermore, we excluded posts exclusively comprising links or hashtags from the dataset. To enable better word comprehension by the models concerning particular words, we adopted a deliberate strategy to standardize slang terms by converting them to their root forms. For instance, "imo" would be changed to "in my opinion", and "lol" to "laughing out loud". This involved curating a dataset encompassing the most frequently employed slang words derived from the Reddit and Twitter social media sites. Subsequently, we chose to retain only specific punctuation marks, namely full stops, commas, exclamation marks and question marks, while eliminating all other punctuations. We made this decision on the belief that these selected punctuations inherently enhanced the model's capacity to acquire a more complex comprehension of sentiment analysis. Furthermore, we applied a uniform transformation to render all words in lowercase, ensuring consistency throughout the dataset. We used the same pre-processing steps for data collected from Twitter, with the addition of an extra step involving the removal of words employed in queries to retrieve relevant tweets.

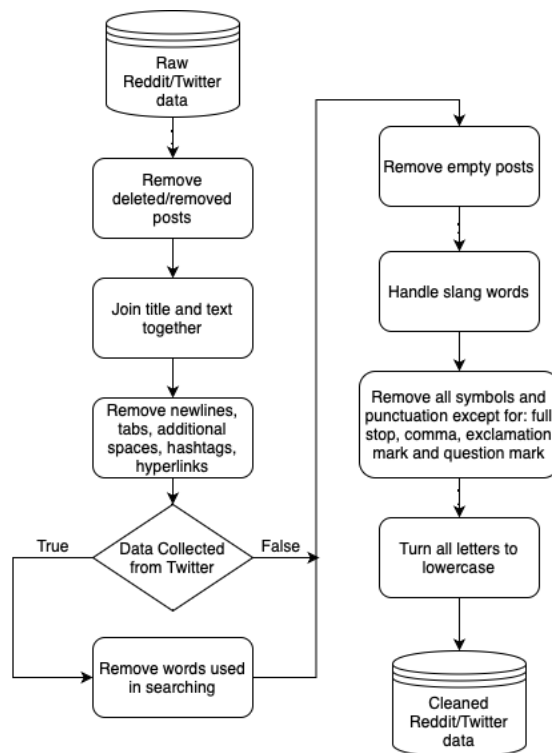


Figure 3.3: Pre-processing steps for Reddit and Twitter data

3.4 Meaningful words detection

We chose several distinct methodologies for the identification of meaningful words:

Supervised learning models: We used the dataset collected from Reddit to train a supervised learning model tailored to predict the classification of Reddit posts into depression/anxiety subreddits or other subreddits. We used exclusively models that incorporated a fitting method providing insight into feature importance: LinearSVC and XGBoost. Upon successful model training, the next step involved extracting the most predictive words from the aforementioned models. These identified terms were regarded as the most meaningful words. Given that feature importance scores were characterized by positive values for the positive class (depression/anxiety) and similarly negative values for the negative class, we based selecting the most meaningful words on the absolute magnitude of these scores derived from the fit method. In our methodological approach, it was imperative to not only identify the most important words but also ascertain the likelihood of those words being masked. This was guided by the normalized weight of the model's fit method assigned to each word. Consequently, words which yielded higher weights in the model's prediction had a higher probability of being masked.

Clustering: Clusters are groups that contain similar entities, thereby encompassing objects sharing similar characteristics. The process of clustering involves the partitioning of objects into these groups. A pivotal rule is to ensure that objects within a given cluster exhibit maximum similarity while objects situated across distinct clusters exhibit dissimilarity. Given the limitations of ML approaches in handling words directly, we used TF-IDF to translate words into representative vectors. Subsequently, KMeans clustering was applied to these representative vectors to create two distinct clusters: one devoted to words related to depression or anxiety and the other encompassing remaining terms. Words were allocated to these clusters based on their affinities with other terms within each cluster. Identifying the most important words relied on noticing the disparities between the centroids, with emphasis placed on attaining maximal differences. We associated the assignment of probabilities for masking with scores derived from this process.

Log-odds: To find the most meaningful words, an initial step involved constructing individual word frequency dictionaries for posts from the depression/anxiety subreddit and those from the random subreddits. These dictionaries contained the respective frequencies of each word occurrence within their respective datasets. Afterwards, we quantified the likelihood of each word appearing within both datasets by calculating probabilities. To establish a comparative measure, we computed odd ratios for each word. This involved dividing the probability of a word's occurrence in the depression/anxiety subreddit by its probability in the random subreddits. Further refinement was achieved by calculating logarithm of the odd ratio for each word. We considered words with a higher log-odds score more important, and their respective log-odds score guided the masking probability.

TF-IDF: we chose two approaches for the implementation of this methodology:

1. **TF-IDF and Chi-Square:** The initial step involved using the TF-IDF method to convert words into vector representations. Subsequently, we used the chi-square test to ascertain whether a word's prevalence within a particular class significantly deviates from what would be anticipated by random chance. Words with a higher chi-square value were deemed more important in distinguishing between the two datasets. These respective chi-square scores guided the determination of masking probabilities.
2. **TF-IDF difference in positive and negative class:** Utilizing the TF-IDF technique, we conducted calculations on both datasets. Subsequently, we computed the disparity between the TF-IDF scores of the depression or anxiety dataset and the random dataset for each word. The selection of the most important words relied upon identifying the words with the highest absolute values of these TF-IDF scores. Concurrently, we determined masking probabilities based on these calculated scores.

Manually selected words: Following the findings presented in Al-Mosaiwi et al. [33], the proportion of absolutist words identified within the test groups for anxiety, depression and suicidal ideation was notably higher compared to the control groups. This observation provided a great opportunity for using these absolutist words as the most meaningful words for the differentiation between posts originating from the depression or anxiety dataset and those from the random dataset. The methodology outlined in the aforementioned paper involved identifying absolutist words through the collaboration of five independent expert judges. Two of the judges were clinical psychologists from the University of Reading Charlie Waller Institute, and three were linguists from the University of Reading School of Clinical Language Science. This process yielded a dictionary encompassing 19 absolutist words. Given the need to assign masking probabilities to these words, we adopted a manual approach, consistently choosing probabilities such as 0.65, 0.5, and 0.35. We applied this uniform approach across all absolutist words.

3.5 Domain Specific Pre-Training

During the phase of DSPT, we implemented selective masking. This involved training models using the distinct set of meaningful word detection strategies identified in the section dedicated to meaningful word detection. We gave these meaningful words a higher probability of being masked compared to other words. To address the inherent constraint posed by the limited vocabulary coverage of the Huggingface built-in tokenizer [16], we adopted an approach of whole word masking [39]. This offered a more efficient and streamlined method for enhancing the capability of neural machine translation models with open-vocabulary translation. The approach achieved this by encoding infrequent and unfamiliar words as sequences of sub-word units allowing for more translations via smaller units than words. To keep the overall masked words in a sentence at 15% similar

to BERT, we lowered the probability of masking non-meaningful words with the following formula [7]:

$$P(W_n) = \frac{\max((|S| \times 0.15) - (|M| \times k), 0)}{|S| \times |M|}$$

Where S is the input sentence, and M are the most meaningful words, k is the average probability of masking these meaningful words.

Algorithm 1 presents the pseudo-code for masking different words, encompassing all the mentioned steps, including ensuring that at least one word was masked in each sentence.

Algorithm 1 Mask Words

```

1: procedure MASKWORDS(tokens)
2:   word_tokens  $\leftarrow$  tokens[non_special_token_mask]
3:   word_token_indices  $\leftarrow$  tokens[non_special_token_mask].indices    ▷ Gets
   indices of non-special tokens
4:   tokens_to_ids  $\leftarrow$  convert_ids_to_tokens(word_tokens)    ▷ Translate these
   tokens to specified words
5:   whole_words  $\leftarrow$  get_whole_words_indices(word_token_indices)    ▷ Gets
   indices of whole words, it contains the same index for a word which is tokenized by
   two subwords
6:   whole_words_set  $\leftarrow$  unique(whole_words)
7:   to_mask  $\leftarrow$  length of whole_words_set  $\times$  0.15
8:   meaningful_word_tokens  $\leftarrow$  tokens of meaningful word from external file
9:   rand  $\leftarrow$  random(length of whole_words_set) ▷ Generate a random number for
   each word
10:  length_meaningful_token  $\leftarrow$  length of most meaningful tokens in word_tokens
11:  list_with_meaningful_tokens  $\leftarrow$  meaningful tokens in tokens
12:  avg_meaningful_word_prob  $\leftarrow$   $\frac{\sum_{i=0}^{\text{length\_meaningful\_token}} \text{probability}(\text{list\_with\_meaningful\_tokens})}{\text{length\_meaningful\_token}}$ 
13:  sum_mng_prob  $\leftarrow$   $\frac{\text{length of } \textit{list\_with\_meaningful\_tokens}}{\text{length of } \textit{list\_with\_meaningful\_tokens}} \times$ 
   avg_meaningful_word_prob
14:  mult_others_meaningful  $\leftarrow$  max((to_mask - sum_mng_prob), 0)
15:  mult_lengths  $\leftarrow$  length of whole_words_set  $\times$  length_meaningful_token
16:  avg_other_word_prob  $\leftarrow$   $\frac{\textit{mult\_others\_meaningful}}{\textit{mult\_lengths}}$ 
17:  if length of rand[rand < 0.15] = 0 then
18:    mask_at_least_one  $\leftarrow$  random index from rand
19:    rand[mask_at_least_one]  $\leftarrow$  0.1    ▷ Make sure that at least one word gets
   masked in each sentence
20:  for (i; i < length of word_tokens; i++) do
21:    word  $\leftarrow$  word_tokens[i]
22:    index  $\leftarrow$  word_token_indices[i]
23:    prob_mask  $\leftarrow$  rand[i]
24:    prob_change  $\leftarrow$  random(0, 1)
25:    if word in meaningful_word_tokens then
26:      prob_for_word  $\leftarrow$  probability of meaningful word from external file
27:    else
28:      prob_for_word  $\leftarrow$  avg_other_word_prob
29:    if prob_mask < prob_for_word then
30:      if prob_change < 0.8 then
31:        tokens[index] = 4
32:      if prob_change  $\geq$  0.9 then
33:        tokens[index] = random word from vocabulary
  return tokens

```

3.6 Fine-Tuning

In the realm of DL, the refinement process known as fine-tuning plays a pivotal role. It entailed the adjustment of pre-trained model parameters tailored to enhance its efficiency within a specific application context. This stage was characterized by a noticeably lower time and computing resource requirement relative to the preceding pre-training phase. The fine-tuning required a dataset of comparably reduced size compared to the initial pre-training phase, albeit reliant upon a labelled dataset. This reliance on labelled data presented challenges in annotations. Particularly valuable when there was an absence of labelled data for the target task, fine-tuning capitalized on the knowledge gleaned during the pre-training phase. We employed a fine-tuning process to calibrate the parameters of the domain-specific pre-trained models. This strategic refinement aimed to enhance the model's proficiency in achieving the intended binary classification outcome, determining whether a given post from Reddit or Twitter exhibited indications of being categorized as either depressive/anxious or not.

Results

This chapter presents the outcomes of employing the aforementioned selective masking strategies using both BERT and RoBERTa as base models for predictions in the depression and anxiety datasets. We generated the results using a Tesla P100 GPU with 16GB of memory. We applied the following metrics:

- **Binary Cross Entropy Loss:** referred to as log loss, this metric measured the difference between the actual binary labels and the predicted probability distribution. The formula for log loss is:

$$BCELoss = -\frac{1}{N} \sum_{i=1}^N y_i(\log(p_i)) + (1 - y_i)(\log(1 - p_i))$$

where N is the number of samples, y_i is the actual label for sample i , and p_i is the predicted probability that sample i belongs to class 1.

- **Accuracy:** accuracy is an evaluation metric which measures how many samples were correctly classified out of all samples. The formula for accuracy is:

$$Accuracy = \frac{TP + FP}{TP + TN + FP + FN}$$

where TP (true positives) are correctly predicted positive samples, TN (true negatives) are correctly predicted negative samples, FP (false positive) is when incorrectly predicting negative samples as positive, and FN (false negatives) is when incorrectly predicting positive samples as negative.

- **Precision:** precision is a metric that measures the accuracy of positive predictions made by a model. In other words, it focused on the model's accuracy when it predicted a positive class. The formula for precision is:

$$Precision = \frac{TP}{TP + FP}$$

- **Recall:** recall measured the ability of a model to correctly identify all relevant instances. It focused on the model’s ability to capture all instances of the positive class. The formula for recall is:

$$Recall = \frac{TP}{TP + FN}$$

- **F1-Score:** f1-Score combined precision and recall into one value. F1-Score is the harmonic mean of precision and recall, and it is especially useful when there are unbalanced class distributions. The formula for f1-score is:

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

In the text below, areas of focus related to depression and anxiety disorders will be referred as the depression and anxiety domains, respectively. Due to memory constraints, we had to limit the training data to a subset. From the pool of gathered posts on Reddit, we specifically chose 24,000 posts for both depression and anxiety domains. The selection process considered the number of likes, serving as a potential filter to identify posts closely associated with the specific subreddit. Additionally, the number of words within each post played a crucial role in the selection process. This was particularly significant for the random class, where the average words per post varied, as outlined in Table 3.1. We gave preference to posts containing more words during the selection process.

4.1 Depression Domain

4.1.1 Domain-Specific Pre-Training

To train the model for the depression domain, we used data from the depression subreddit and random posts from the most popular subreddits. We configured the model parameters with the following specifications: a learning rate of 0.00002, a batch size of 8, a maximum of 512 words in a sentence, a weight decay of 0.01, and trained for 3 and 5 epochs. For selective masking, we implemented all previously mentioned masking strategies with two distinct settings for mask probabilities. In the first setting, we used a min-max scaler to scale weights from 0.6 to 0.16, focusing on the top 1000 words. In the second setting, we used the min-max scaler to scale weights from 0.5 to 0.2, focusing on the top 500 words. Both BERT and RoBERTa served as base models for DSPT, resulting in the training of a total of 50 models for depression. We used tokenizers 'bert-base-cased' and 'roberta-base' in the process.

Masking Strategy	Epochs	Min-Max MP	Loss Score		Time (s)
			Train	Validation	
Absolutist Words	3	0.50-0.50	0.135	0.127	5816.750
Cluster	3	0.16-0.60	0.124	0.116	8350.200
LinearSVC	3	0.16-0.60	0.131	0.123	8005.710
Log-Odds	3	0.16-0.60	0.136	0.128	6464.430
TF-IDF	3	0.16-0.60	0.118	0.110	9172.140
TF-IDF Neg-Pos	3	0.16-0.60	0.105	0.098	11941.300
XGBoost	3	0.16-0.60	0.131	0.124	7327.540
None	3	0.15-0.15	0.135	0.128	5415.870

Table 4.1: DSPT results for depression domain using BERT as base model

Table 4.1 displays the results obtained when using BERT as a base model for DSPT with data from Reddit. Only the outcomes with the best f1-score in the evaluation dataset for each masking strategy were presented, while all other results are available in the appendix. The findings indicated that employing TF-IDF difference in positive and negative class as a masking strategy enhances the model’s proficiency in identifying masked words. Moreover, it was evident that selective masking, which involved different masking probabilities for different words, outperformed a uniform mask probability for all words across all proposed masking strategies.

The observed difference between the validation and training scores could be attributed to the limited training epochs, set at 3 or 5, which might have been insufficient for the model to over-fit and yield a higher training score than the validation score. In the context of Large Language Models (LLMs), an increase in the number of epochs typically led to a continual increase in the training accuracy, while the validation accuracy stabilized after a certain amount of epochs. The lower training accuracy might have also indicated the utilization of random weight initialization. Moreover, in the early training epochs, the model was learning from the training data, yet it might not have fully converged, resulting in a lower training accuracy than the validation accuracy.

It was evident that certain masking strategies required more time for training compared to others. Specifically, TF-IDF Neg-Pos generally demanded more time, while training with absolutist words was less time-consuming. This discrepancy was primarily influenced by the number of words involved in each strategy, as absolutist words masked fewer words than other masking strategies. Notably, in the choice of either the top 1000 or 500 words with the highest weight from masking strategies, the number of words that were masked differed depending on the number of words that were part of the tokenizer vocabulary.

Table 4.2 details the number of words which were part of BERT and RoBERTa tokenizer vocabulary for each masking strategy. Notably, strategies like TF-IDF had a higher count of recognized words, indicating a preference for commonly used words for this masking strategy which were part of tokenizer vocabularies. In contrast, masking strategies like log-odds tended to select words more associated with medications for depression, which effectively distinguished between depression and random domains, albeit these words were usually not part of the tokenizer’s vocabulary. Absolutist words masking strategy contained only 19 words with all of them being part of tokenizer vocabulary. Another noteworthy observation was that RoBERTa tokenizer identified a greater number of

4. RESULTS

words compared to BERT tokenizer due to its more extensive vocabulary; RoBERTa tokenizer had a vocabulary of 50,265 words, while BERT tokenizer had a vocabulary of 28,996 words. Since the objective of this thesis was to focus solely on the most important words without including common words, we did not fix the number of words masked for each strategy. Only the words among the top 1000 and top 500 most important words, which were part of the tokenizer vocabulary, were masked with a higher probability.

Masking Strategy	Top 500 Words		Top 1000 Words	
	BERT	RoBERTa	BERT	RoBERTa
Absolutist Words	19*	19*	19*	19*
Cluster	478	492	811	896
LinearSVC	327	381	661	779
Log-Odds	82	111	198	259
TF-IDF	443	457	872	906
TF-IDF Neg-Pos	492	497	989	997
XGBoost	420	464	828	921

Note: Absolutist Words strategy contained only 19 words

Table 4.2: Number of words which were present in tokenizer for each masking strategy for depression domain

Table 4.3 displays the results obtained from using RoBERTa as a base model for DSPT using data from Reddit. Similar to BERT, only the results with the best f1-score in the evaluation dataset for each masking strategy were presented. RoBERTa outperformed BERT in correctly predicting the masked words for most strategies, with the best-observed performance when employing the TF-IDF Neg-Pos masking strategy. The pre-training time for both BERT and RoBERTa was approximately the same. However, TF-IDF Neg-Pos took more time to train due to the higher number of words which were part of RoBERTa tokenizer vocabulary.

Overall, a higher probability for masking important words improved the validation score in most masking strategies during DSPT compared to using a uniform fixed percentage for masking. Furthermore, when RoBERTa served as the base model, the validation score was consistently better across all masking strategies compared to using BERT as the base model.

Masking Strategy	Epochs	Min-Max MP	Loss Score		Time (s)
			Train	Validation	
Absolutist Words	5	0.65-0.65	0.086	0.083	9852.740
Cluster	5	0.16-0.60	0.080	0.073	11513.230
LinearSVC	3	0.20-0.50	0.086	0.080	6950.180
Log-Odds	3	0.16-0.60	0.090	0.083	5962.890
TF-IDF	5	0.16-0.60	0.074	0.071	15084.920
TF-IDF Neg-Pos	3	0.20-0.50	0.064	0.060	10219.490
XGBoost	3	0.20-0.50	0.089	0.083	6712.250
None	3	0.15-0.15	0.092	0.086	5710.280

Table 4.3: DSPT results for depression domain using RoBERTa as base model

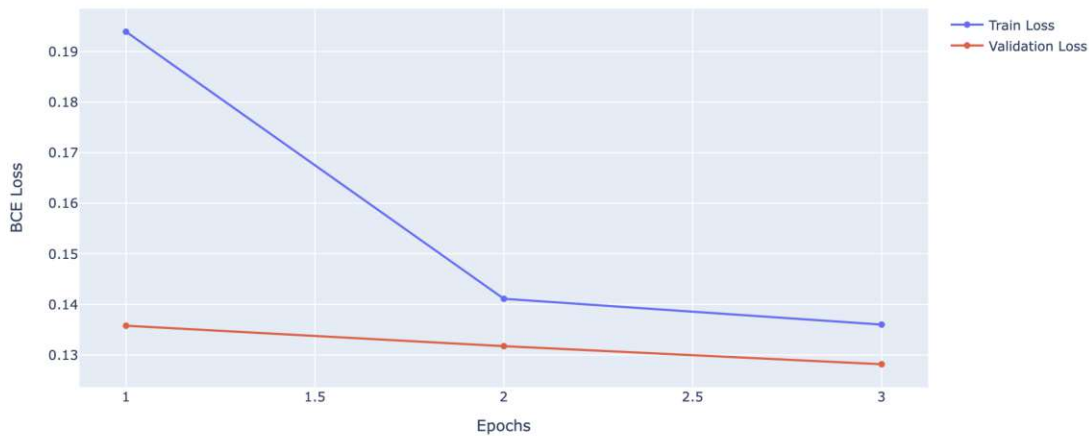


Figure 4.1: Binary Cross Entropy Loss for Log-Odds strategy

When we employed both BERT and RoBERTa as a base model, the validation loss was lower than the training loss. Figure 4.1 displays the results after each epoch when using BERT as a base model and using log-odds as a masking strategy. It illustrated that the training loss was notably higher, particularly in the initial epoch, as the model was in the process of learning from the training data and had not yet fully converged. After the initial epoch, the training and validation loss progressively approached each other. After running an additional 2 epochs of training, 5 in total, it was noted that the validation loss continued to be lower than the training loss. Given this observation and a minimal reduction of loss with further training and the constraints imposed by GPU memory limitations, it became impractical to train the model using a higher number of epochs.

4.1.2 Fine-Tuning and Evaluation

To fine-tune the model in the depression domain, we used data gathered from Reddit, whereas we used data from Twitter to evaluate the results. We configured the model parameters as follows: a learning rate of 0.00002, a batch size of 8, a maximum of 512 words in a sentence, a weight decay of 0.01, and trained for 3 epochs. The model structure is presented in Figure 4.2, where the top section features the pre-trained model. This was followed by a dropout layer designed to avoid overfitting, succeeded by a linear layer incorporating a Gaussian Error Linear Unit (GELU) activation function. The choice of the GELU activation function over Rectified Linear Unit (ReLU) was because GELU weights inputs based on their percentiles rather than their sign. This characteristic enabled GELU to accommodate small negative values when the input was negative, thereby providing a richer gradient for back-propagation.

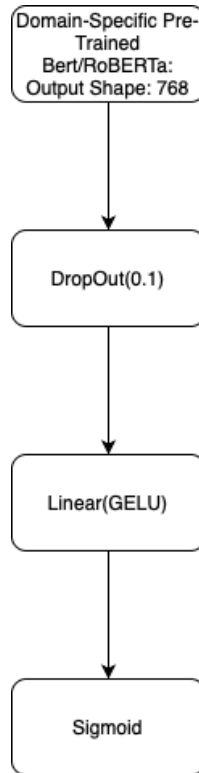


Figure 4.2: DSPT model structure

To assess the effectiveness of different masking strategies, we maintained a straightforward model structure with fixed parameters across all experiments. We chose these parameters following recommendations from the authors of the BERT paper [11].

Masking Strategy	Reddit Dataset		Twitter Dataset		
	Train Accuracy	Validation Accuracy	Precision	Recall	F1-Score
Absolutist Words	0.987	0.973	0.830	0.798	0.814
Cluster	0.987	0.981	0.780	0.858	0.817
LinearSVC	0.985	0.978	0.843	0.781	0.811
Log-Odds	0.986	0.978	0.837	0.733	0.782
TF-IDF	0.986	0.978	0.803	0.770	0.786
TF-IDF Neg-Pos	0.987	0.973	0.760	0.877	0.815
XGBoost	0.984	0.971	0.841	0.735	0.784
None	0.985	0.971	0.832	0.836	0.834
Plain BERT	0.981	0.966	0.820	0.589	0.686
MentalBERT	0.972	0.956	0.803	0.535	0.642

Table 4.4: Fine-Tuning best results for depression domain using BERT as a base model for all masking strategies

As presented in Table 4.4, the model effectively captured the data structure from Reddit, demonstrating high accuracy on both the training and validation sets. The

best-performing model, in terms of validation accuracy, was achieved when we used the clustering masking strategy to identify the most meaningful words that had a higher probability of getting masked. Overall, all models had a decent performance. Compared to other state-of-the-art models like BERT and MentalBERT, our selective masking strategies outperformed plain BERT and MentalBERT across all masking strategies.

Upon examining the evaluation scores, the best-performing model when using BERT as the base model, based on f1-score, emerged when we pre-trained the model on domain data with uniform probabilities for masking each word. This model achieved an f1-score of 0.834, a recall of 0.836, and a precision of 0.831. Analyzing recall, which indicated the model’s ability to correctly detect posts annotated as depressive, TF-IDF Neg-Pos outperformed other masking strategies. This was reasonable, considering that this model had the highest validation score in finding the masked words in the DSPT phase, with this model having a recall score of 0.877. However, it incurred more false positives, resulting in a lower precision score. Regarding precision, the model with the best performance used linearSVC as a masking strategy, achieving a precision score of 0.842.

Upon comparing the results with plain BERT and MentalBERT, it became evident that all models performed better in the evaluation set. This difference was particularly notable regarding recall score, where plain BERT and MentalBERT struggled with many false negatives. This was due to the fact that BERT was trained on general data, lacking specific knowledge about the specific patterns in the depression domain, while MentalBERT, being pre-trained on a broad spectrum of mental health problems, proved too generalized for the nuances of the depression domain.

Masking Strategy	Reddit Dataset		Twitter Dataset		
	Train Accuracy	Validation Accuracy	Precision	Recall	F1-Score
Absolutist Words	0.985	0.978	0.778	0.795	0.786
Cluster	0.983	0.978	0.779	0.892	0.832
LinearSVC	0.983	0.977	0.775	0.863	0.817
Log-Odds	0.985	0.979	0.824	0.859	0.841
TF-IDF	0.985	0.981	0.799	0.854	0.826
TF-IDF Neg-Pos	0.982	0.981	0.770	0.758	0.764
XGBoost	0.984	0.984	0.811	0.798	0.805
None	0.982	0.914	0.857	0.467	0.605
Plain RoBERTa	0.983	0.981	0.800	0.707	0.750
MentalRoBERTa	0.985	0.954	0.853	0.530	0.654

Table 4.5: Fine-Tuning best results for depression domain using RoBERTa as a base model for all masking strategies

Examining the outcomes obtained when using RoBERTa as the base model in Table 4.5, all models yielded promising results in the Reddit data, with XGBoost masking strategy emerging as the top performer. Looking at the evaluation scores derived from Twitter data, the model employing the log-odds masking strategy demonstrated the best performance, achieving an f1-score of 0.841. The best-performing model for recall was when clustering was used as a masking strategy with a score of 0.892. Precision-wise, the most effective masking strategy was when using XGBoost, with a score of

4. RESULTS

0.857. The comparison between RoBERTa and BERT as base models in the evaluation dataset revealed no significant difference, possibly attributed to the limited number of epochs which we used for pre-training the models. Overall, all models achieved a good performance, particularly regarding recall, when compared to the state-of-the-art models BERT and MentalRoBERTa.

Masking Strategy	Recall		Precision		F1-Score	
	Mean	Std	Mean	Std	Mean	Std
XGBoost	0.723	0.084	0.832	0.040	0.770	0.031
Cluster	0.774	0.082	0.790	0.022	0.780	0.037
LinearSVC	0.707	0.117	0.821	0.035	0.754	0.063
Absolutist Words	0.683	0.131	0.808	0.038	0.732	0.070
Log-Odds	0.668	0.111	0.805	0.031	0.727	0.073
TF-IDF	0.619	0.213	0.820	0.031	0.682	0.161
TF-IDF Neg-Pos	0.557	0.232	0.812	0.052	0.631	0.156

Table 4.6: Average scores and standard deviation for each Masking Strategy over all trained models for depression domain

Table 4.6 illustrates that, when calculating the mean f1-scores and their standard deviation from the mean for each model for each masking strategy, XGBoost had the best results. It was closely followed by a cluster masking strategy, indicating that these models exhibited greater stability than the other models. Additionally, the table revealed a similar average score and standard deviation for precision across all models, signifying a high level of stability in terms of precision. The distinguishing factor relied upon recall scores, where certain masking strategies demonstrated a better ability to detect tweets belonging to the depression class.

4.2 Anxiety Domain

4.2.1 Domain-Specific Pre-Training

We followed a similar methodology as the one employed for the depression domain to train models for anxiety, using the following parameters: a learning rate of 0.00002, a batch size of 8, a maximum of 512 words in a sentence, a weight decay of 0.01, and training for 3 and 5 epochs.

For the anxiety domain, we applied the same masking probabilities as those used for the depression domain. In the first setting, the masking probability ranged from 0.6 to 0.16, focusing on the top 1000 most important words, while in the second setting, it ranged from 0.5 to 0.2, focusing on the top 500 words. Table 4.7 shows the best-performing model for each masking strategy. When using BERT as a base model, the model achieving the best results in terms of identifying masked words was when using TF-IDF Neg-Pos, closely followed by TF-IDF and clustering as masking strategies. The results for each strategy are presented in the appendix, revealing that increasing the number of epochs during the pre-training phase contributed to improved performance in identifying the masked words in all scenarios.

Masking Strategy	Epochs	Min-Max MP	Loss Score		Time (s)
			Train	Validation	
Absolutist Words	3	0.65-0.65	0.111	0.105	5706.060
Cluster	3	0.16-0.60	0.102	0.097	8343.030
LinearSVC	3	0.16-0.60	0.104	0.098	6946.510
Log-Odds	5	0.16-0.60	0.105	0.101	8888.560
TF-IDF	5	0.16-0.60	0.095	0.090	11987.110
TF-IDF Neg-Pos	5	0.16-0.60	0.079	0.076	14931.920
XGBoost	3	0.16-0.60	0.108	0.102	6684.870
None	3	0.15-0.15	0.111	0.105	5131.950

Table 4.7: DSPT results for anxiety domain using BERT as base model

Similarly, as for the depression domain, Table 4.8 displays the count of words which were present in tokenizer vocabularies for each masking strategy. Notably, the TF-IDF Neg-Pos masking strategy had the highest number of words, which were also present in the tokenizer vocabulary, while log-odds had the lowest number of words. The difference arose because log-odds assigned higher importance to words associated with anxiety medications, which were not present in the tokenizer vocabulary, resulting in a lower count of recognised words by the tokenizer. It is worth noting that for absolutist words strategy according to the findings presented in Al-Mosaiwi et al. [33] the number of absolutist words identified by the judges was 19 with all being present in the tokenizer vocabulary. Comparing the number of words selected by masking strategies present in tokenizer vocabulary for anxiety and depression domain we can notice that the number of words present in tokenizer vocabulary was lower for anxiety domain. This arose because the anxiety domain had a higher number of words associated with anxiety medications which were not present in the tokenizer vocabulary.

Masking Strategy	Top 500 Words		Top 1000 Words	
	BERT	RoBERTa	BERT	RoBERTa
Absolutist Words	19*	19*	19*	19*
Cluster	451	473	798	865
LinearSVC	321	379	679	794
Log-Odds	79	103	176	233
TF-IDF	445	462	868	905
TF-IDF Neg-Pos	487	494	982	993
XGBoost	409	459	845	928

Note: Absolutist Words strategy contained only 19 words.

Table 4.8: Nr. of words which were present in tokenizer for each masking strategy for anxiety domain

Examining Table 4.9, it was evident that when employing RoBERTa as the base model, the best performing model was once again when using TF-IDF Neg-Pos masking strategy, attributed to the same reasons observed in the depression domain. Following closely were TF-IDF and clustering masking strategy. Notably, when we used RoBERTa as a base model, better performance was achieved in identifying masked words compared to when using BERT as a base model. Generally, all masking strategies showed an increase in validation accuracy compared to pre-training with a fixed uniform percentage for all words in DSPT. The only exception was when we used the log-odds masking strategy,

4. RESULTS

which might be attributed to the small number of words that were masked with a higher probability.

Masking Strategy	Epochs	Min-Max MP	Loss Score		Time (s)
			Train	Validation	
Absolutist Words	3	0.50-0.50	0.074	0.070	5568.010
Cluster	3	0.16-0.60	0.069	0.064	7775.50
LinearSVC	3	0.16-0.60	0.069	0.065	7440.810
Log-Odds	3	0.20-0.50	0.074	0.070	6030.630
TF-IDF	3	0.20-0.50	0.067	0.062	7708.680
TF-IDF Neg-Pos	5	0.16-0.60	0.054	0.053	16469.390
XGBoost	3	0.16-0.60	0.072	0.068	6772.430
None	3	0.15-0.15	0.074	0.070	5949.400

Table 4.9: DSPT results for anxiety domain using RoBERTa as base model

Figure 4.3 presents the train and validation loss when using RoBERTa as a base model with log-odds masking strategy. It reveals that validation loss was lower than training loss, with the scores converging more closely in subsequent epochs. Notably, in the initial epoch, the difference between training and validation loss was significantly more pronounced than when using BERT as a base model. This observation suggested that RoBERTa takes longer to converge when compared to BERT, potentially due to its larger and more complex parameter structure.

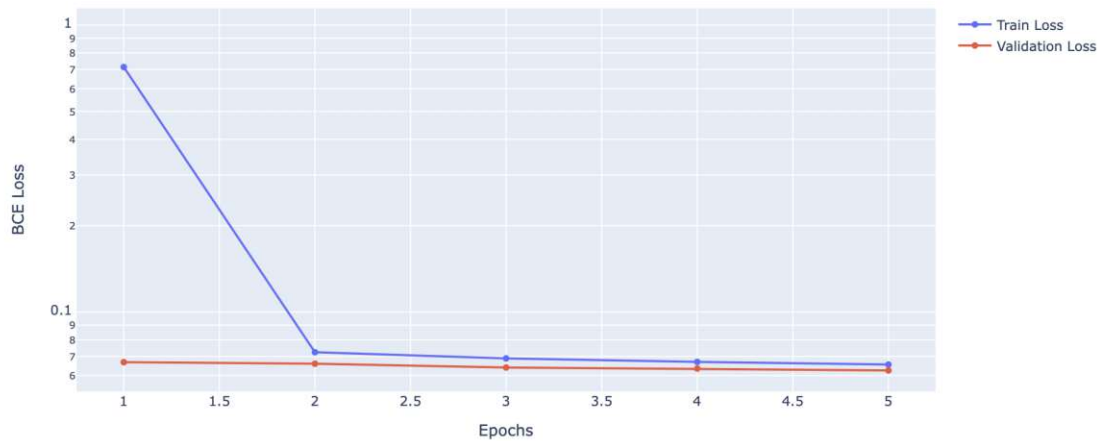


Figure 4.3: Binary Cross Entropy Loss for Log-Odds strategy

4.2.2 Fine-Tuning and Evaluation

We trained the models using the following parameters: a learning rate of 0.00002, a batch size of 8, a maximum of 512 words in a sentence, a weight decay of 0.01, and training for 3 epochs. The model structure was similar to that used in the depression domain and is illustrated in Figure 4.2. When we employed Reddit data and BERT as a base model in the evaluation dataset, presented in Table 4.10, the best performing model was when using TF-IDF Neg-Pos. The remaining strategies achieved comparable

performances, except for MentalBERT and when we used fixed uniform masking. In the case of the latter, overfitting occurred during DSPT in the last epoch, resulting in a diminished performance in the validation dataset. Upon reviewing the evaluation results, the best-performing model in terms of f1-score was when using absolutist words as a masking strategy, with a probability of 65% of masking more meaningful words, closely followed by TF-IDF Neg-Pos. Absolutist words also demonstrated a better performance in terms of recall, with a score of 0.974. Assessing precision, when we used the XGBoost masking strategy during DSPT it stood out as the best-performing model with a score of 0.962. It is noteworthy that in terms of recall, all masking strategies had a better performance compared to both plain BERT and MentalBERT. However this performance was not replicated in precision, which indicates that the models were able to detect more instances from the anxiety class albeit with a higher rate of errors. In contrast, plain BERT demonstrated a high level of certainty in its predictions for anxiety class while missing numerous posts labeled as belonging to the anxiety class. Looking into f1-scores, all masking strategies, excluding tf-idf, outperformed plain BERT, fixed uniform masking and MentalBERT. This suggested that selective masking enhanced the model’s ability to learn more about the patterns of the anxiety domain.

Masking Strategy	Reddit Dataset		Twitter Dataset		
	Train Accuracy	Validation Accuracy	Precision	Recall	F1-Score
Absolutist Words	0.985	0.974	0.889	0.974	0.930
Cluster	0.985	0.977	0.893	0.956	0.924
LinearSVC	0.987	0.976	0.893	0.952	0.922
Log-Odds	0.987	0.977	0.890	0.890	0.890
TF-IDF	0.983	0.972	0.847	0.897	0.871
TF-IDF Neg-Pos	0.987	0.978	0.923	0.930	0.927
XGBoost	0.986	0.978	0.946	0.897	0.921
None	0.988	0.926	0.962	0.563	0.710
Plain BERT	0.985	0.974	0.902	0.879	0.890
MentalBERT	0.953	0.933	0.933	0.662	0.774

Table 4.10: Fine-Tuning best results for anxiety domain using BERT as a base model for all masking strategies

In Table 4.11, it can be seen that MentalRoBERTa achieved the highest validation accuracy, closely followed by XGBoost, with all models demonstrating decent performance. Upon analyzing the evaluation results, TF-IDF Neg-Pos was the best performing model, followed by the absolutist words masking strategy. TF-IDF Neg-Pos also outperformed other models regarding recall, while TF-IDF had the best precision score. Overall, all models displayed better performance compared to plain RoBERTa, MentalRoBERTa, and uniform masking. This once again underscored the role of selective masking in enhancing the model’s understanding of domain-specific patterns. When comparing the outcomes obtained by using BERT and RoBERTa as base models, we can see improvement across all masking strategies in terms of validation accuracy. However, this improvement is not reflected in the evaluation dataset. This suggests that when using BERT as a base model, the models did a better job at generalizing on different social media platforms.

4. RESULTS

Masking Strategy	Reddit Dataset		Twitter Dataset		
	Train Accuracy	Validation Accuracy	Precision	Recall	F1-Score
Absolutist Words	0.985	0.979	0.893	0.956	0.924
Cluster	0.987	0.979	0.934	0.886	0.909
LinearSVC	0.987	0.978	0.840	0.949	0.891
Log-Odds	0.985	0.978	0.916	0.926	0.921
TF-IDF	0.986	0.978	0.951	0.860	0.903
TF-IDF Neg-Pos	0.984	0.972	0.884	0.978	0.928
XGBoost	0.983	0.979	0.933	0.875	0.903
None	0.984	0.978	0.879	0.879	0.879
Plain RoBERTa	0.982	0.975	0.844	0.956	0.897
MentalRoBERTa	0.990	0.981	0.852	0.868	0.860

Table 4.11: Fine-Tuning best results for anxiety domain using RoBERTa as a base model for all masking strategies

Masking Strategy	Recall		Precision		F1-Score	
	Mean	Std	Mean	Std	Mean	Std
Cluster	0.922	0.026	0.878	0.041	0.898	0.019
XGBoost	0.863	0.091	0.912	0.037	0.883	0.045
TF-IDF Neg-Pos	0.899	0.120	0.904	0.041	0.895	0.053
Log-Odds	0.819	0.090	0.896	0.043	0.854	0.059
Absolutist Words	0.865	0.128	0.911	0.042	0.880	0.067
TF-IDF	0.776	0.207	0.905	0.057	0.817	0.127
LinearSVC	0.714	0.368	0.722	0.356	0.714	0.356

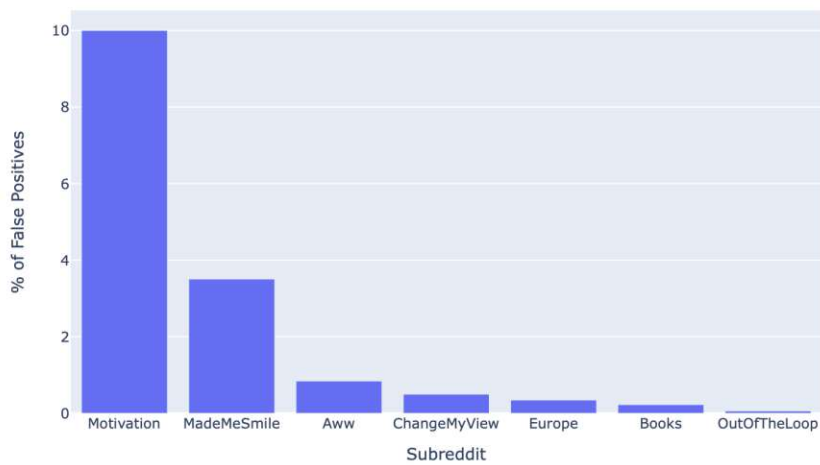
Table 4.12: Average scores and standard deviation for each Masking Strategy over all trained models for anxiety domain

Table 4.12 highlights the stability of masking strategies; cluster masking strategy had the highest stability, followed closely by XGBoost, as indicated by their low standard deviations. As in the depression domain, the variability in scores was predominantly driven by the recall score, showing a significantly higher standard deviation compared to precision scores. However, it is worth noting that linearSVC encountered overfitting in one of the trained models during the last epoch, which contributed to its higher standard deviation.

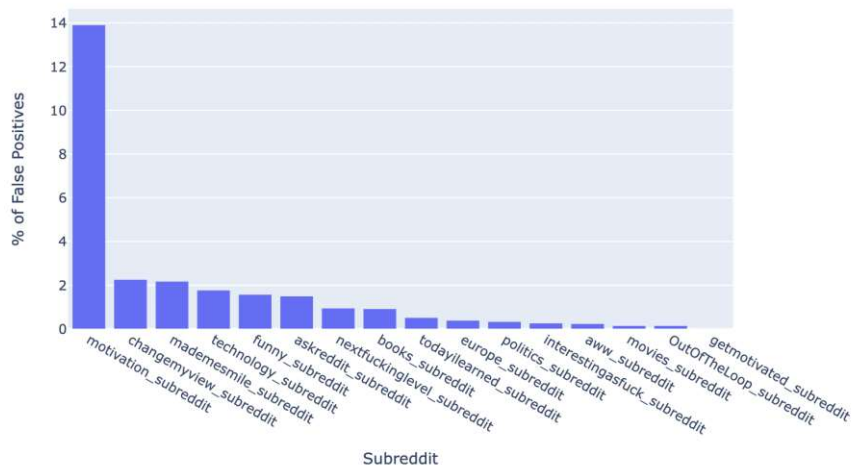
4.3 Comparative Analysis and Explainability

Selective masking demonstrated better performance in both anxiety and depression domains. The outcomes were particularly notable for the anxiety disorder, showing improved performance with selective masking and when using plain RoBERTa. This discrepancy between the depression and anxiety domains might be attributed to the nature of training data sourced from anxiety and depression subreddits. It is worth noting that the rules of the r/Depression subreddit instructed users to submit posts related to suicidal thoughts or feelings on r/Suicidewatch instead. This aspect posed a limitation for the depression model as it might not have captured all the patterns associated with depression due to it being trained only using data from r/Depression. Figure 4.4(a) illustrates the percentage of posts from various subreddits which were classified into the depression class. Notably, around 10% of posts collected from r/Motivation

were classified to the depression class. This aligned with expectations, considering that motivational content was also on the r/Depression subreddit. The second highest false positive rate occurred in r/MadeMeSmile, a subreddit known for positive content, which could occasionally overlap with content found in r/Depression. This pattern was similarly reflected in Figure 4.4(b), illustrating the percentage of posts from various subreddits which were classified to the anxiety class. R/Motivation stood out with the highest rate of false positives, while other subreddits exhibited lower percentages.



((a)) False positives for depression model



((b)) False positives for anxiety model

Figure 4.4: Percentage of false positives in different subreddits

We used both LIME and SHAP to enhance the interpretability of the models. LIME works by generating a dataset of similar posts by perturbing the features of the original post, this dataset is created by sampling from a distribution centered around the original post. LIME then fits an interpretable model like ridge regression to the generated dataset. Afterwards, LIME analyzes the coefficients of the ridge regression and identifies the most important features for the prediction of the original instance. Whereas SHAP first computes the Shapley values, which are based on cooperative game theory. Shapley values distribute the contribution of each feature to the prediction by considering all possible combinations of features and their marginal contributions.

Given the complexity of interpreting over 50 models individually, we adopted a selection approach. The chosen strategy was based on identifying the masking strategy that yielded the best average score relative to its standard deviation. According to Tables 4.6 and 4.12, the selected strategy with the lowest standard deviation for the depression domain was using XGBoost, whereas for the anxiety domain, was clustering masking strategy. For the depression domain, RoBERTa served as the base model, pre-trained for three epochs, with a min-max masking probability of 0.2-0.5. In contrast, for anxiety, BERT served as the base model, pre-trained for three epochs, with a min-max masking probability of 0.16-0.6. To enhance interpretability further, we provided both global and local explanations.

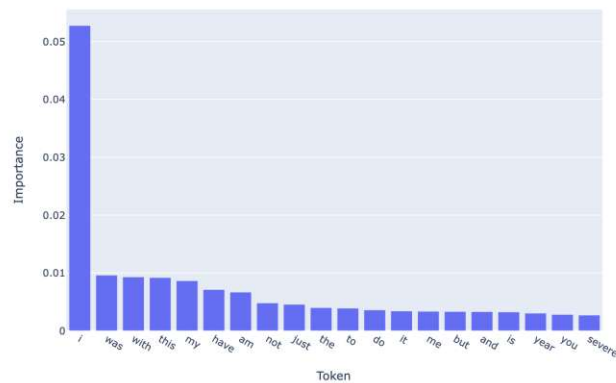
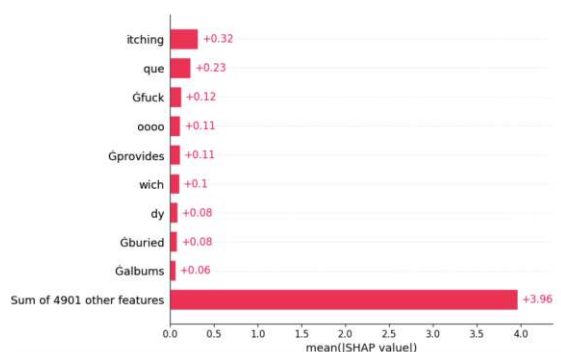
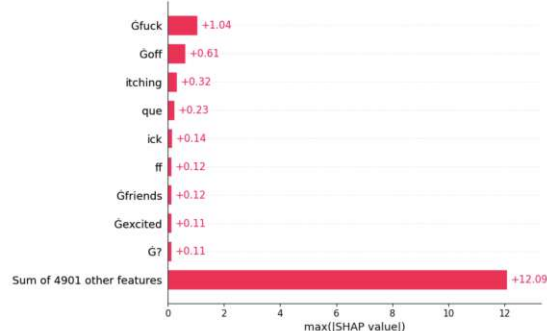


Figure 4.5: Global explanation when using LIME for depression domain

Figure 4.5 displays the global explanations derived from LIME for the depression domain. Due to the technical limitations of the LIME package in Python, we implemented a workaround to approximate global explanations. Instead of a direct global explanation, we generated a local explanation for each instance in the test data, recorded the weights of individual words present in the tokenizer vocabulary, and computed the average weight for each word. Words with the highest average weight were presented. The figure displays the significance of first-person pronouns, which consistently carried substantial weight and played a crucial role in the model's predictions. Auxiliary verbs and frequently used words like "severe" in the depression subreddit also appeared as important contributors.



(a) Mean score for word



(b) Max score for word

Figure 4.6: Global explanation when using SHAP for depression domain

Figure 4.6 presents the global explanations using SHAP, and it consists of two subplots: Figure 4.6(a) displaying the average weights of words, and Figure 4.6(b) showing the maximum weights of these words. We generated these visualizations using the SHAP library in Python. We used RobertaTokenizer to explain the results. Words starting with "G" represented whole words identified by the tokenizer, while other entries signified subwords. The analysis revealed a notable average score for curse words, followed by specific verbs that were instrumental in identifying tweets belonging to the random class. In Figure 4.6(b), the examination of maximum scores highlighted the importance of curse words. Furthermore, punctuation, such as question marks, played a significant role in influencing the model's predictions.

We applied the same approach employed for the depression model to the anxiety model for extracting global explanations using LIME, as displayed in Figure 4.7. Similar to the depression model, first-person pronouns played a crucial role in the anxiety model. However, in addition to first-person pronouns, words associated with expressing feelings played a higher significance. Words linked to anxiety, such as "mental", "severe", "panic" and "attacks" carried a notable weight. It is worth noting that the anxiety model demonstrated a stronger focus on the positive class compared to the depression model, which could contribute to its better performance.

4. RESULTS

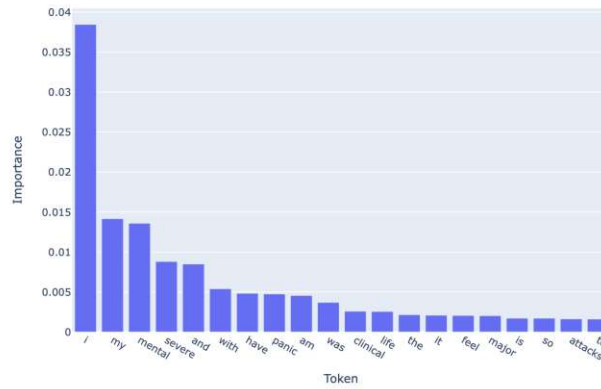
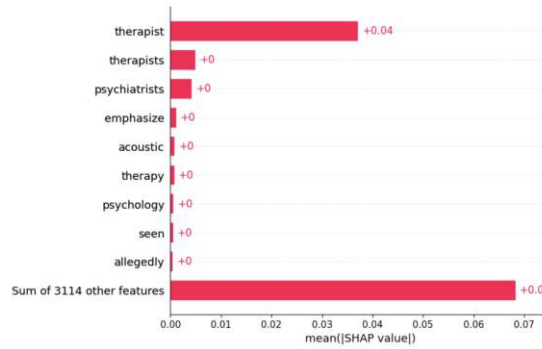
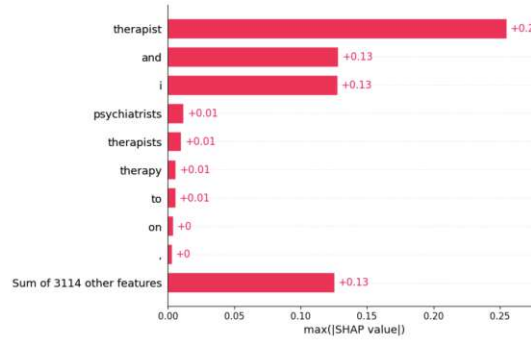


Figure 4.7: Global explanation when using LIME for anxiety domain



((a)) Mean score for word



((b)) Max score for word

Figure 4.8: Global explanation when using SHAP for anxiety domain

Figure 4.8 illustrates the global explanations for the anxiety domain using SHAP. Similar to the depression domain in Figure 4.8(a), the average weight of words was displayed, whereas in Figure 4.8(b), the maximum weight of words was shown. Notably, in the anxiety domain, whole words held more significance compared to subwords, with words associated

with the positive class consistently displaying high average scores. The maximum scores revealed that in addition to words linked to anxiety, first-person pronouns and punctuation also played an essential role in the model's predictions.

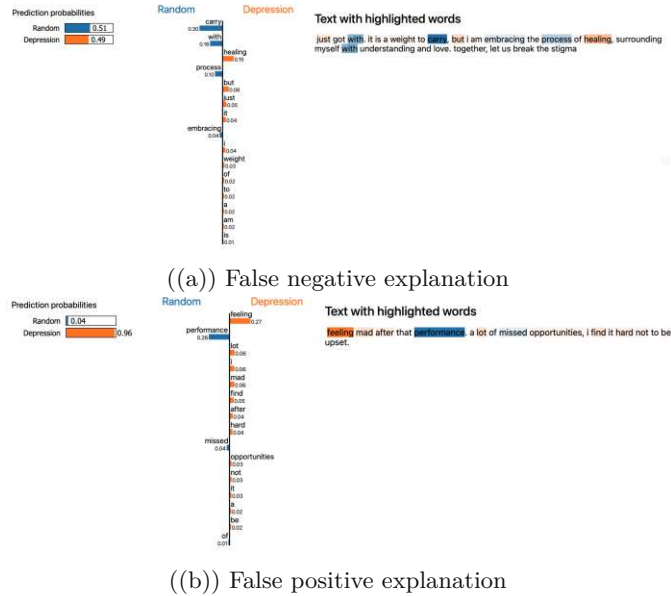


Figure 4.9: Local explanation when using LIME for depression domain

Given the importance of sharing a specific post for local explanations, we opted to leverage ChatGPT to generate tweets associated with depression, anxiety and random tweets. Subsequently, we made predictions on these generated tweets and chose two tweets, one representing a false positive (misclassified as belonging to the depression or anxiety class when it didn't) and another a false negative (misclassified as belonging to the random class when it didn't).

We selected the following tweets for the depression domain:

1. **False Negative:** Just got diagnosed with depression. It is a weight to carry, but I'm embracing the process of healing, surrounding myself with understanding and love. Together, let us break the stigma. #MentalHealth #Stigma
2. **False Positive:** Feeling mad after that performance. A lot of missed opportunities, I find it hard not to be upset. #PremierLeague #Chelsea

In Figure 4.9, the local explanations using LIME for both false negative and false positive instances were displayed. In the case of false negative, as indicated in Figure 4.9(a), on the left, the model predicted with a low certainty that this tweet belongs to the random class. In the middle of the figure, words and their contributions to the prediction were

displayed, where words such as "healing", "I", and "am" carried weights that drove the prediction towards the depression class. Words like "carry", "with", and "process" were associated with the random class, and their higher weights led the model to predict this tweet as belonging to the random class. In the right of the figure, the tweet was displayed after performing pre-processing steps, with each word highlighted according to its weight and the class it was associated with. Given that all models displayed a lower recall score than a precision score, which indicated a tendency for many false negatives, this example represented such cases. Figure 4.9(b) explains the model's prediction for a false positive. The model was confident of its prediction, displaying a tendency to be more confident when predicting for the depression class than the random class. The analysis revealed that the model was influenced by words related to feelings and first-person pronouns in classifying this tweet as belonging to the depression class, with most words contributing to the depression class. We generated these local explanation plots using LIME library in Python.

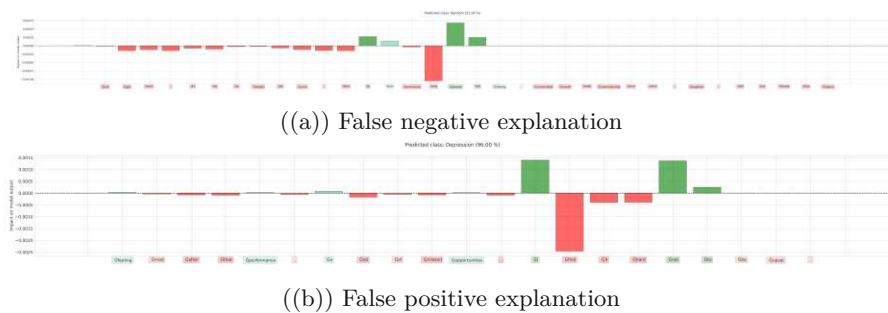


Figure 4.10: Local explanation when using SHAP for depression domain

In Figure 4.10, the local explanations for the same tweets for the depression domain using SHAP are presented. We generated these visualizations by following the methodology outlined in the "BERT meets Shapley" paper [40], which provided clearer insight compared to the SHAP library. The SHAP values explained how the impact of unmasking each word changed the models' output from where the entire input was masked to the final prediction value.

In Figure 4.10(a), the false negative example is displayed, where most words exhibited a negative weight, indicating their association with the random class. Similar to the findings when we used LIME, first-person pronouns tended to be more related to the depression class. However, we observed a distinction where the word "healing" had a lower weight, and "process" had a positive connotation. This divergence in explanation might have arisen from the different algorithms employed by LIME and SHAP for generating explanations, with LIME relying on a local surrogate model, while SHAP was based on Shapley values and cooperative game theory.

In Figure 4.10(b), the local explanation for the false positive example for the depression domain is presented. It revealed that words used for expressing feelings, first-person

pronouns, and words expressing negative sentiment contributed to the depression class. When compared with LIME explanations for the same tweet, differences became apparent, particularly in certain words like "hard", "mad", and "performance". These differences could be attributed to the distinct algorithms employed by LIME and SHAP for generating their explanations.

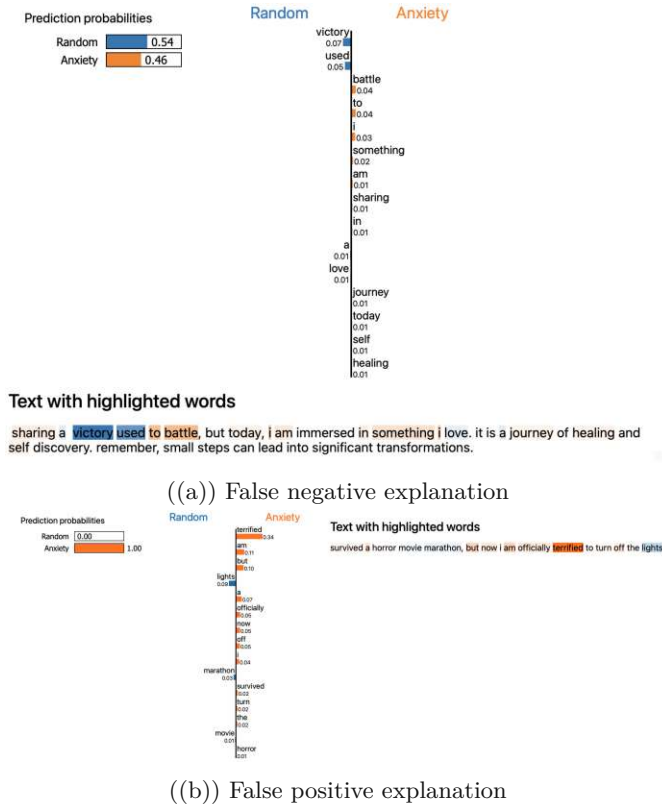


Figure 4.11: Local explanation when using LIME for anxiety domain

We selected the following tweets for the anxiety domain:

1. **False Negative:** Sharing a victory used to battle anxiety, but today, Im immersed in something I love. It is a journey of healing and self discovery. Remember, small steps can lead into significant transformations. #MentalHealth #Victory #Healing
2. **False Positive:** Survived a horror movie marathon, but now I am officially terrified to turn off the lights. #HorrorMovie #Marathon

Similarly, for the anxiety domain, we generated local explanations using LIME and SHAP. Figure 4.11 illustrates the local explanations when employing LIME for the previously mentioned tweets. In Figure 4.11(a), the false negative example for the anxiety domain

4. RESULTS

is displayed. It revealed that positive words like "victory" and "love" contributed towards the random class, while words such as "battle", "I", and "healing" contributed to the anxiety class.

In Figure 4.11(b), the local explanation using LIME for the false positive example is presented. The analysis revealed that words such as "terrified", "I", and "survived" significantly influenced the model's prediction towards the anxiety class. Furthermore, it was shown that the model had a high level of certainty when predicting for the anxiety class compared to when predicting for the random class.

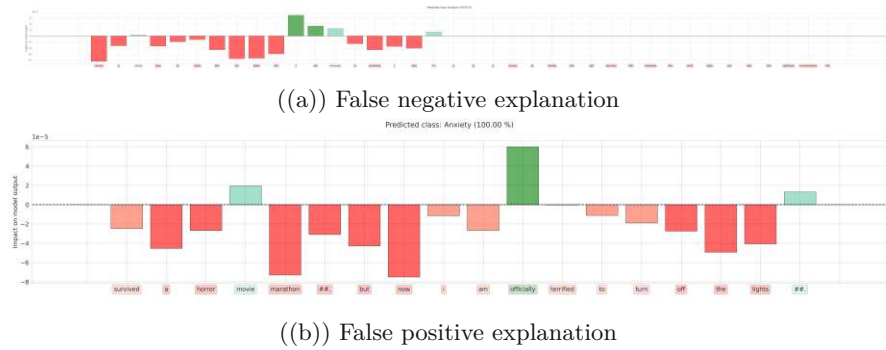


Figure 4.12: Local explanation when using SHAP for anxiety domain

In Figure 4.12, the local explanations using SHAP for the anxiety domain are displayed. Figure 4.12(a) presents the false negative explanations, highlighting that most words contributed towards the random class, except first-person pronouns and words like "immersed," which appeared less frequently in Reddit posts. In Figure 4.12(b), the false positive example revealed that only words such as "officially", "movie", and punctuations contributed towards the anxiety class. We observed notable differences when comparing these explanations with LIME explanations, where words like "I", "terrified", and "survived" contributed more towards the random class. These variances might have appeared from algorithmic differences between LIME and SHAP.

Conclusion

In this chapter, we explore key findings and elaborate on our contributions to the current state-of-the-art models.

5.1 Main Conclusions

In this master's thesis, we suggested using selective masking for DSPT, assigning a higher masking probability to more meaningful words as opposed to other words. We propose various methods to identify these meaningful words and their corresponding probability, including:

1. Supervised learning models
2. Clustering
3. TF-IDF
4. Log-Odds
5. Manually selected words

For the baseline model, we chose either BERT or RoBERTa depending on which model we used as a base model in DSPT. We trained the models using Reddit data while using Twitter data for the model evaluation. We implemented pre-processing steps to clean both Reddit and Twitter data, including removing words used for searching tweets, expanding abbreviations, and removing hashtags, links and mentions from tweets. We generated labelling using distant supervision, with each post labelled based on the subreddit to which it was posted.

Due to GPU memory constraints, we trained the models using only a subset of Reddit data. The selection of this subset was determined by factors such as the number of words per post and the number of likes per post, which indicated that the chosen posts were more closely associated with the given subreddit.

We chose various configurations during DSPT, resulting in the training of 50 models for both depression and anxiety domains. We used metrics including accuracy, precision, recall and f1-score to assess the performance of models. Notably, even with plain BERT and plain RoBERTa, the models demonstrated decent performance, particularly in terms of precision, where they were good at spotting the negative class. However, the baseline models faced challenges with detecting the positive class. Additionally, we compared our results against other DSPT models like MentalBERT and MentalRoBERTa, and observed that these models also exhibited a lower recall score.

To enhance the recall score, we adjusted the probability assigned to words with greater significance in depression and anxiety disorders. Using our proposed methods for identifying these words and determining their probabilities, we observed that, during DSPT with these strategies for selective masking, most trained models improved recall scores compared to the baseline models. This suggested that, throughout DSPT, the models effectively learned patterns specific to these domains. As we trained different models with various configurations for each masking strategy, our focus was on masking strategies that not only produced good results but also demonstrated stability (lower standard deviation). Notably, better and more consistent results were achieved when using XGBoost and clustering as masking strategies.

We arrive at the following answers to the research questions raised throughout our work:

1. How does pre-training with domain-specific unlabeled data influence the model's performance?

Pre-training with domain-specific unlabeled data using the flat masking strategy did not yield significantly improved results compared to plain BERT or plain RoBERTa. This observation could be attributed to the low number of epochs and the relatively small dataset used during the pre-training phase, making the plain masking strategy less effective.

2. How does selective masking of words specifically linked to depression or anxiety improve the model's performance?

Selective masking notably improved the model's performance, particularly when employing XGBoost and clustering masking strategy, displaying improvements in the recall score. However, it is important to note that this approach had a drawback: it was computationally more expensive compared to the straightforward use of BERT and RoBERTa.

3. How well does the model generalize when using Reddit data for training and Twitter data for evaluation?

We observed good results in the evaluation dataset utilizing Twitter data when using selective masking. Notably, the performance was strong in the anxiety disorder domain, where the difference between the results from the validation dataset, which used data from Reddit and the results from the evaluation dataset, which used data from Twitter, was not substantial.

5.2 Contribution to the state-of-the-art

This thesis introduced a model applicable to various social media platforms through cross-platform evaluations, producing good results in the evaluation dataset. Specifically, when using selective masking with XGBoost and clustering masking strategies, Table 5.1 illustrates a substantial improvement in recall compared to baseline models and other state-of-the-art models. The results presented in this table were based on the settings that yielded the best results for these masking strategies. Notably, we observed a significant increase in recall for the depression domain. Additionally, we noticed a comparable recall score in the anxiety domain, whether using plain RoBERTa or using the clustering masking strategy with BERT as the base model, with the latter also showing an improved precision.

Since existing state-of-the-art models were less transparent and challenging to interpret due to being black-box models, this thesis enhanced interpretability by providing global and local explanations. This improvement in interpretability increased the transparency of the model.

Domain	Selective masking strategy	Evaluation Recall
Anxiety	Plain Bert	0.879
Anxiety	Plain RoBERTa	0.956
Anxiety	MentalBERT	0.662
Anxiety	MentalRoBERTa	0.868
Anxiety	Cluster/ BERT	0.956
Depression	Plain Bert	0.686
Depression	Plain RoBERTa	0.750
Depression	MentalBERT	0.642
Depression	MentalRoBERTa	0.654
Depression	Cluster/ RoBERTa	0.892

Table 5.1: Recall scores for baseline models and best stable models

APPENDIX **A**

Appendix

A.1 Figures

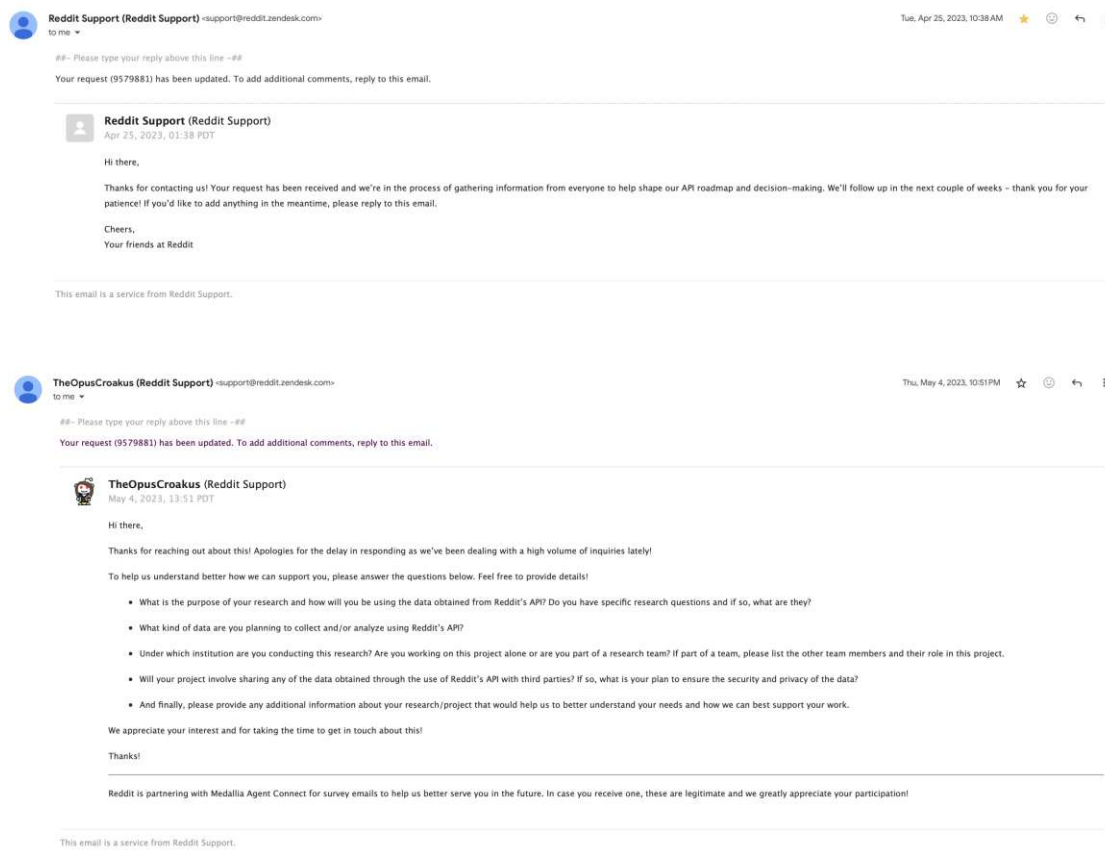


Figure A.1: First 2 replies from Reddit admins regarding Reddit data usage

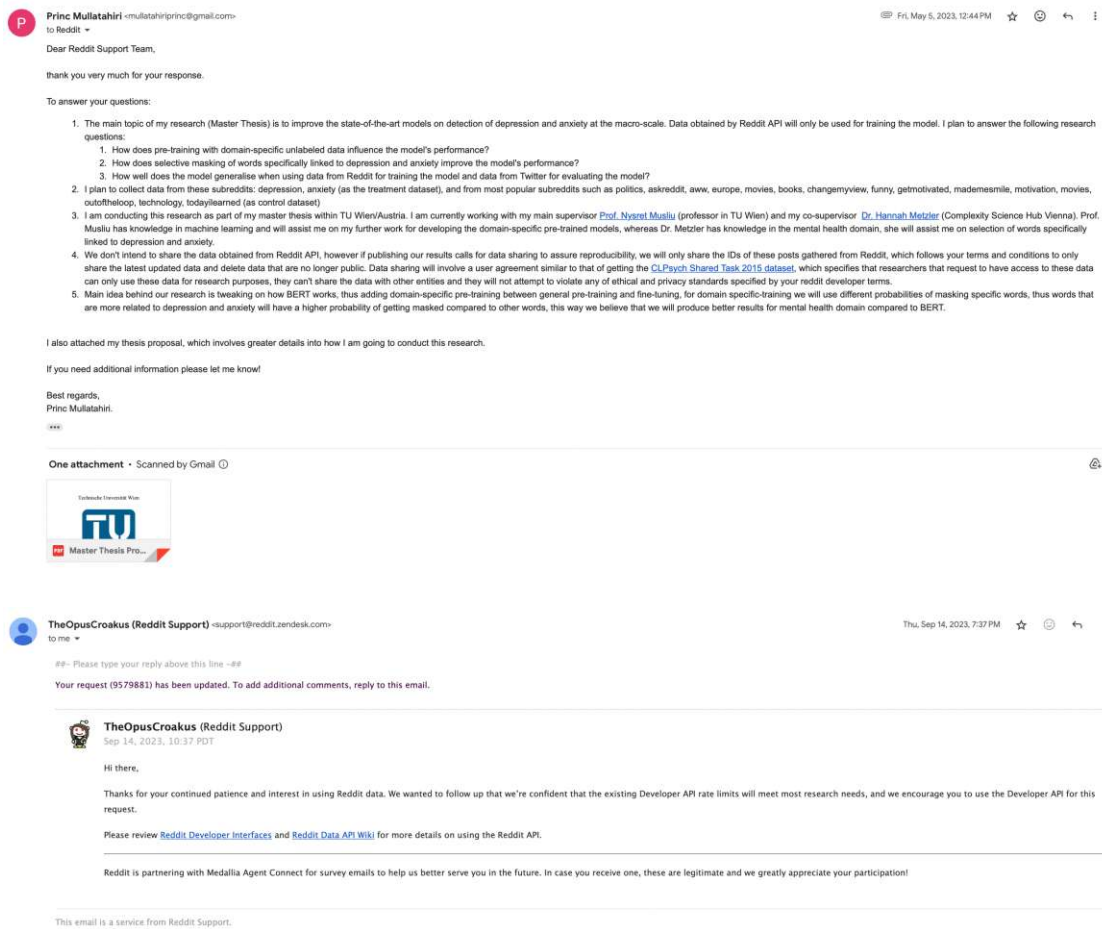


Figure A.1: Last 2 replies from Reddit admins regarding Reddit data usage

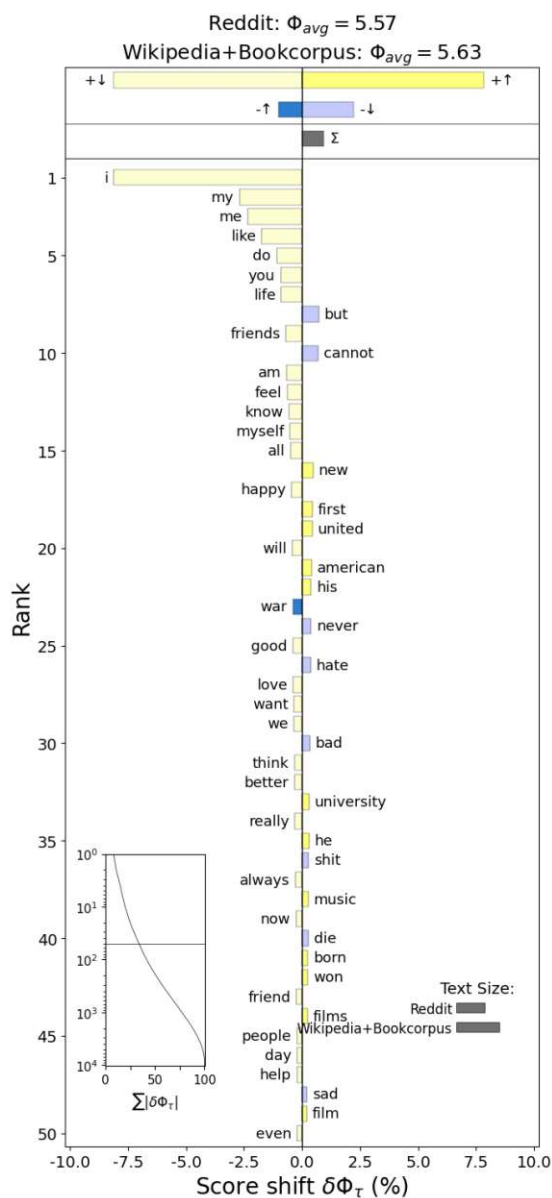


Figure A.2: Dictionary-based sentiment analysis for Reddit and Wikipedia+Bookcorpus for depression domain

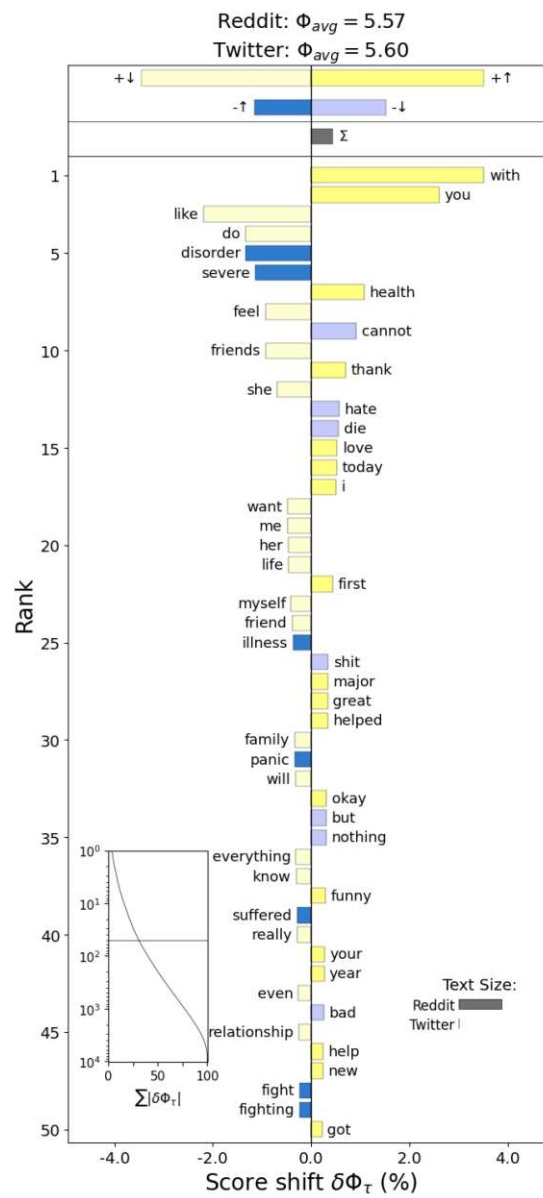


Figure A.3: Dictionary-based sentiment analysis for Reddit and Twitter data for depression domain

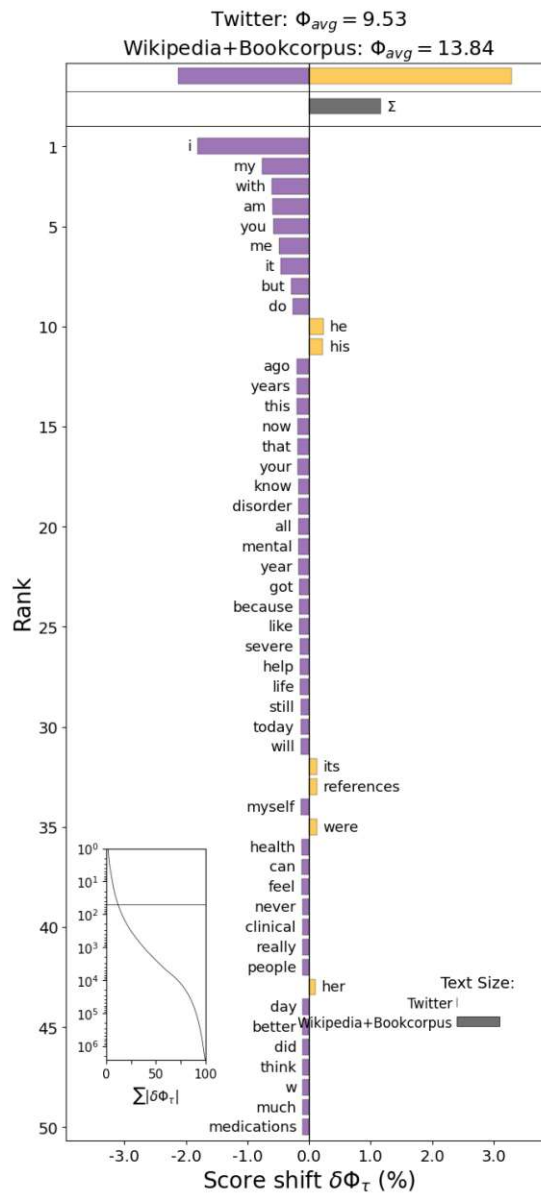


Figure A.4: Shannon Entropy Shifts between Twitter and Wikipedia+Bookcorpus for depression domain

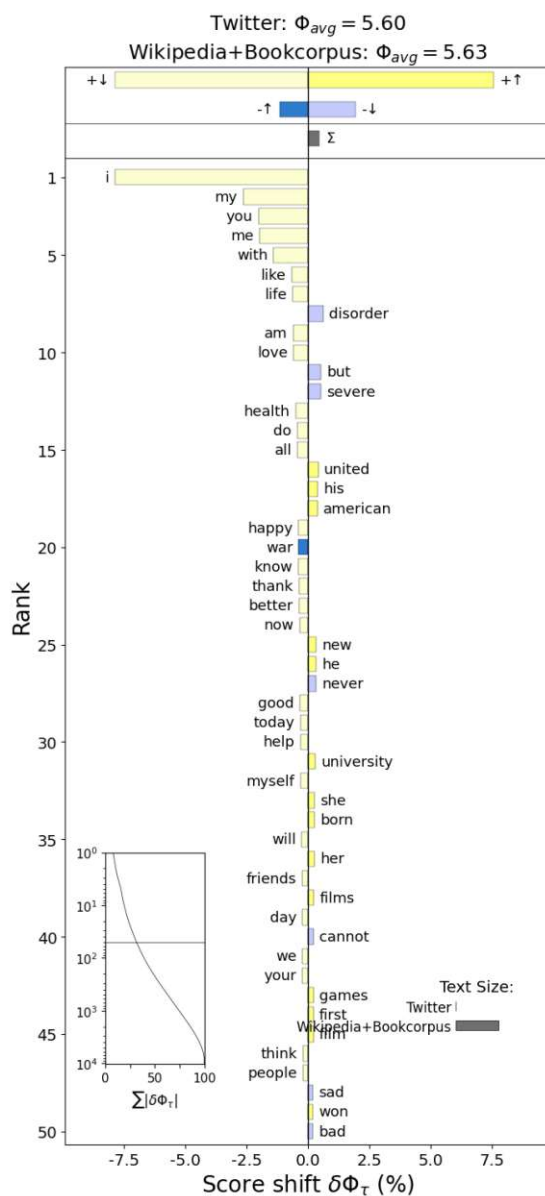


Figure A.5: Dictionary-based sentiment analysis for Twitter and Wikipedia+Bookcorpus for depression domain

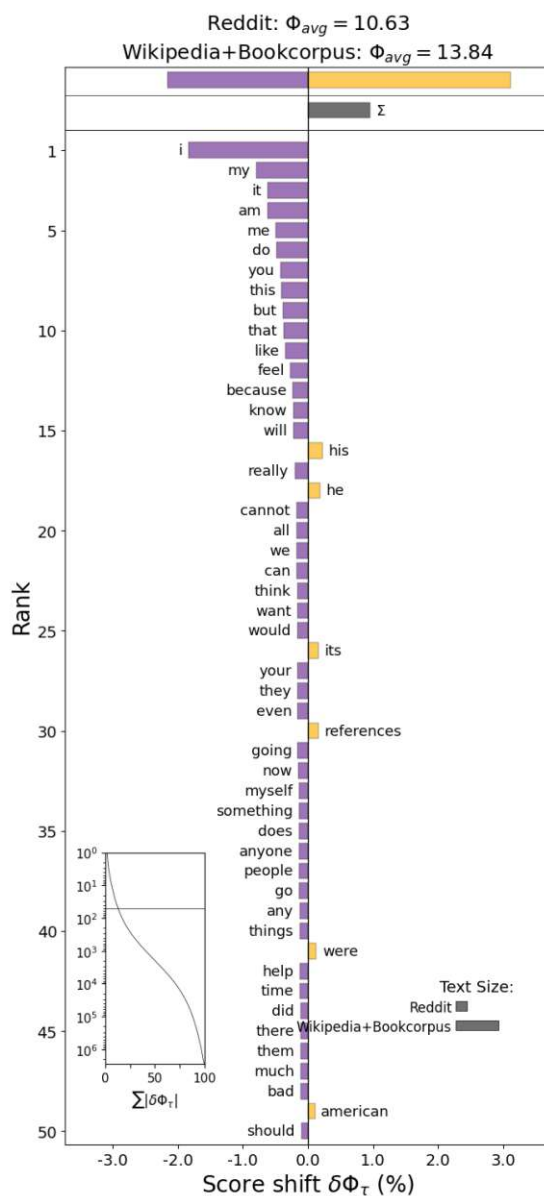


Figure A.6: Shannon Entropy Shifts between Reddit and Wikipedia+Bookcorpus for anxiety domain

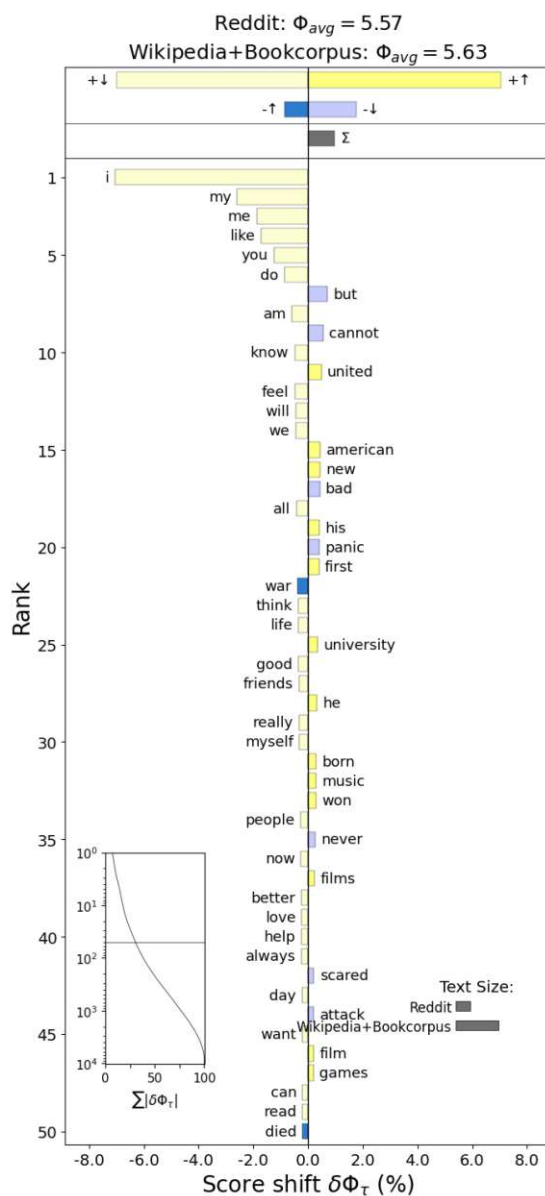


Figure A.7: Dictionary-based sentiment analysis for Reddit and Wikipedia+Bookcorpus for anxiety domain

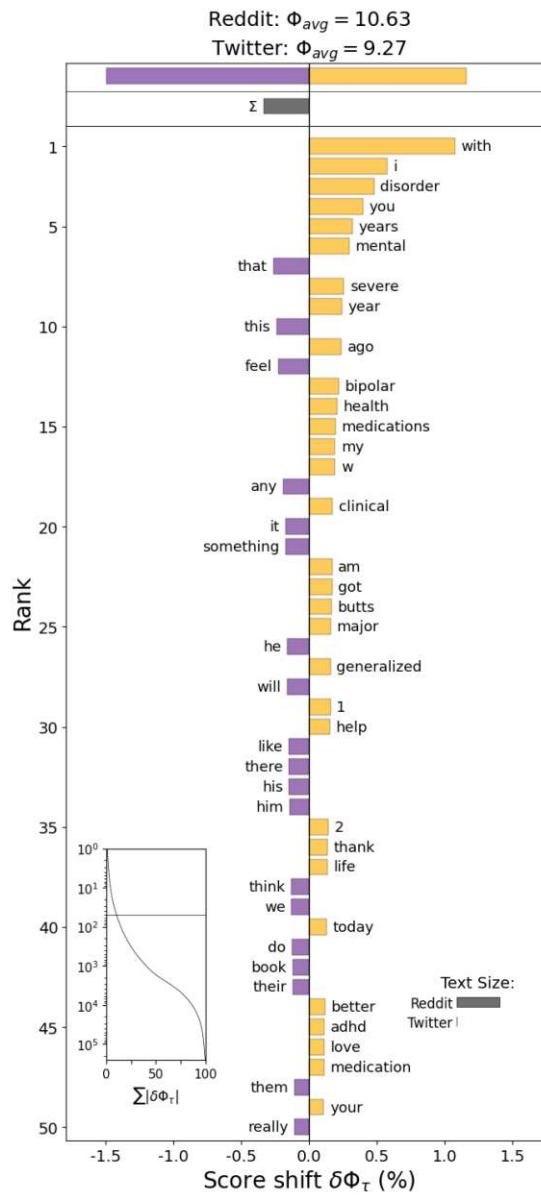


Figure A.8: Shannon Entropy Shifts between Reddit and Twitter data for anxiety domain

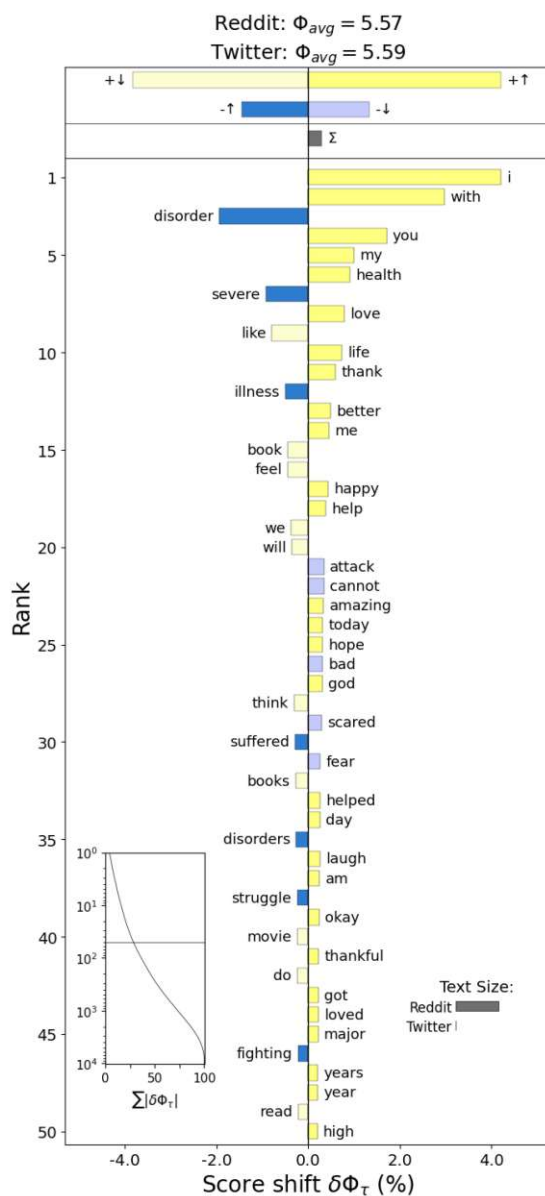


Figure A.9: Dictionary-based sentiment analysis for Reddit and Twitter data for anxiety domain

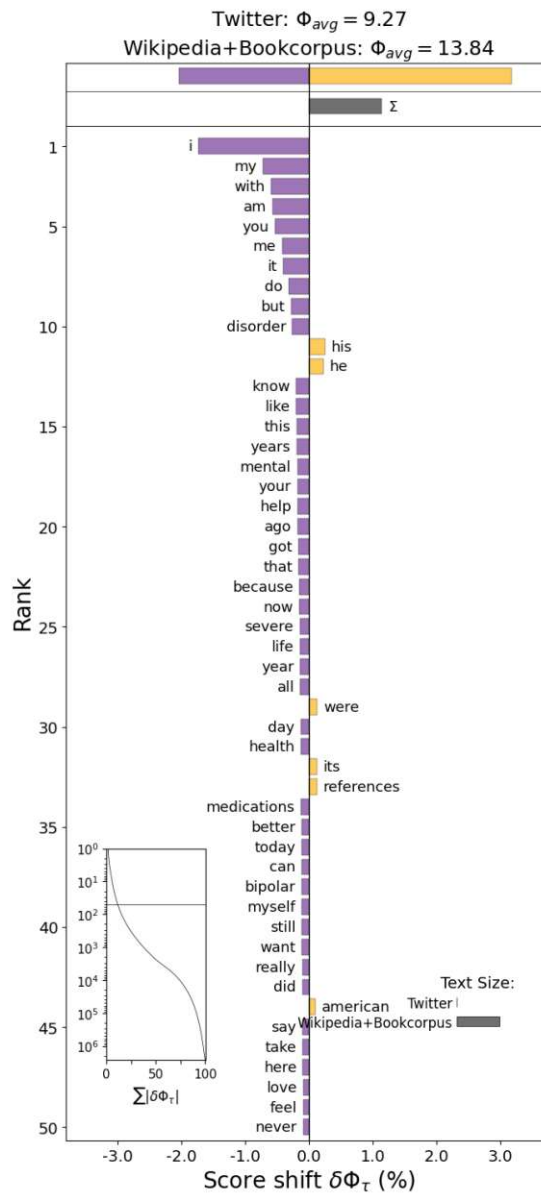


Figure A.10: Shannon Entropy Shifts between Twitter and Wikipedia+Bookcorpus for anxiety domain

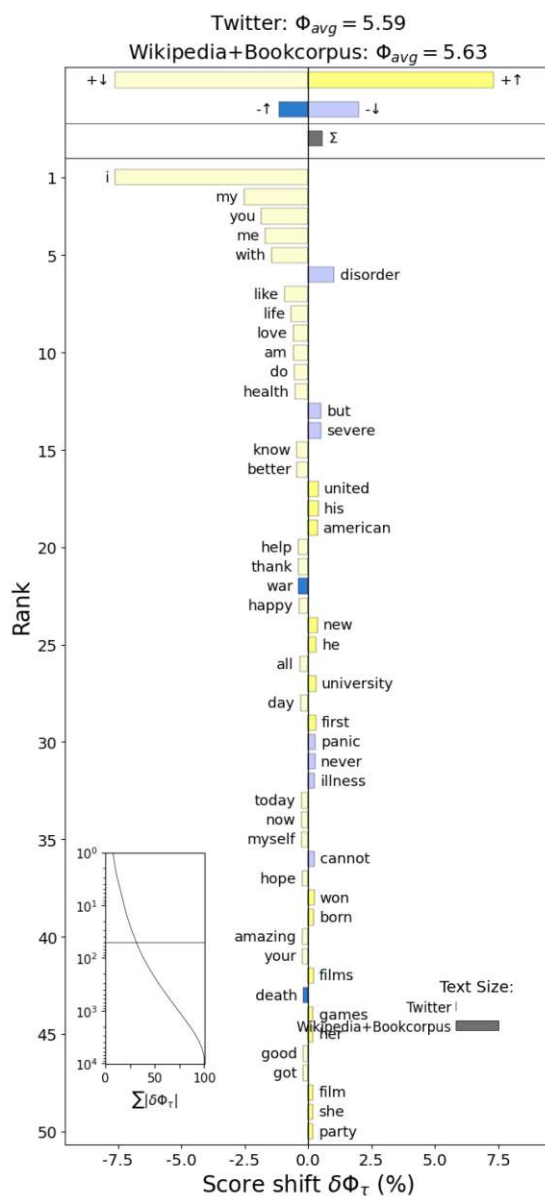


Figure A.11: Dictionary-based sentiment analysis for Twitter and Wikipedia+Bookcorpus for anxiety domain

A.2 Tables

Masking Strategy	Epochs	Min-Max MP	Loss Score		Time (s)
			Train	Validation	
Absolutist Words	3.0	0.35-0.35	0.135	0.129	5881.910
Absolutist Words	3.0	0.50-0.50	0.135	0.128	5816.750
Absolutist Words	3.0	0.65-0.65	0.135	0.127	5839.270
Cluster	3.0	0.16-0.60	0.124	0.117	8350.200
LinearSVC	3.0	0.16-0.60	0.131	0.123	8005.710
Log-Odds	3.0	0.16-0.60	0.136	0.128	6464.430
TF-IDF	3.0	0.16-0.60	0.118	0.111	9172.140
TF-IDF Neg-Pos	3.0	0.16-0.60	0.105	0.098	11941.300
XGBoost	3.0	0.16-0.60	0.132	0.124	7327.540
Absolutist Words	5.0	0.35-0.35	0.129	0.122	9827.670
Absolutist Words	5.0	0.50-0.50	0.128	0.122	9725.010
Absolutist Words	5.0	0.65-0.65	0.127	0.122	10135.280
Cluster	5.0	0.16-0.60	0.117	0.112	14117.430
LinearSVC	5.0	0.16-0.60	0.125	0.120	12707.780
Log-Odds	5.0	0.16-0.60	0.129	0.123	10938.480
TF-IDF	5.0	0.16-0.60	0.112	0.106	16811.340
TF-IDF Neg-Pos	5.0	0.16-0.60	0.098	0.093	20145.940
XGBoost	5.0	0.16-0.60	0.125	0.120	12560.740
Cluster	3.0	0.20-0.50	0.124	0.117	8309.960
LinearSVC	3.0	0.20-0.50	0.130	0.122	7044.550
Log-Odds	3.0	0.20-0.50	0.136	0.128	5925.730
TF-IDF	3.0	0.20-0.50	0.119	0.112	9044.780
TF-IDF Neg-Pos	3.0	0.20-0.50	0.100	0.093	10335.820
XGBoost	3.0	0.20-0.50	0.132	0.125	7245.870
None	3.0	0.15-0.15	0.136	0.128	5415.870

Table A.1: DSPT all results for depression domain using BERT as base model

Masking Strategy	Epochs	Min-Max MP	Loss Score		Time (s)
			Train	Validation	
Absolutist Words	3.0	0.35-0.35	0.090	0.084	5641.040
Absolutist Words	3.0	0.50-0.50	0.090	0.084	5622.700
Absolutist Words	3.0	0.65-0.65	0.094	0.087	6213.710
Cluster	3.0	0.16-0.60	0.083	0.077	7940.290
LinearSVC	3.0	0.16-0.60	0.087	0.081	6758.850
Log-Odds	3.0	0.16-0.60	0.090	0.084	5962.890
TF-IDF	3.0	0.16-0.60	0.078	0.073	8671.940
TF-IDF Neg-Pos	3.0	0.16-0.60	0.068	0.064	10392.620
XGBoost	3.0	0.16-0.60	0.088	0.082	7160.170
Absolutist Words	5.0	0.35-0.35	0.087	0.084	9553.270
Absolutist Words	5.0	0.50-0.50	0.087	0.084	9360.750
Absolutist Words	5.0	0.65-0.65	0.086	0.084	9852.740
Cluster	5.0	0.16-0.60	0.080	0.073	11513.230
LinearSVC	5.0	0.16-0.60	0.084	0.081	11669.080
Log-Odds	5.0	0.16-0.60	0.088	0.084	10403.900
TF-IDF	5.0	0.16-0.60	0.074	0.072	15084.920
TF-IDF Neg-Pos	5.0	0.16-0.60	0.066	0.063	18949.690
XGBoost	5.0	0.16-0.60	0.085	0.082	13098.410
Cluster	3.0	0.20-0.50	0.082	0.077	8246.330
LinearSVC	3.0	0.20-0.50	0.087	0.081	6950.180
Log-Odds	3.0	0.20-0.50	0.090	0.084	6711.960
TF-IDF	3.0	0.20-0.50	0.079	0.074	8509.700
TF-IDF Neg-Pos	3.0	0.20-0.50	0.064	0.061	10219.490
XGBoost	3.0	0.20-0.50	0.090	0.084	6712.250
None	3.0	0.15-0.15	0.093	0.087	5710.280

Table A.2: DSPT all results for depression domain using RoBERTa as base model

Masking Strategy	Epochs	Min-Max MP	Loss Score		Time (s)
			Train	Validation	
Absolutist Words	3.0	0.35-0.35	0.111	0.105	5606.150
Absolutist Words	3.0	0.50-0.50	0.111	0.111	5749.340
Absolutist Words	3.0	0.65-0.65	0.111	0.105	5706.060
Cluster	3.0	0.16-0.60	0.102	0.097	8343.030
LinearSVC	3.0	0.16-0.60	0.104	0.098	6946.680
Log-Odds	3.0	0.16-0.60	0.111	0.106	5877.670
TF-IDF	3.0	0.16-0.60	0.100	0.094	8030.980
TF-IDF Neg-Pos	3.0	0.16-0.60	0.085	0.081	10573.610
XGBoost	3.0	0.16-0.60	0.108	0.102	6684.870
Absolutist Words	5.0	0.35-0.35	0.105	0.101	8712.790
Absolutist Words	5.0	0.50-0.50	0.105	0.101	8676.880
Absolutist Words	5.0	0.65-0.65	0.105	0.100	8758.700
Cluster	5.0	0.16-0.60	0.097	0.093	11886.920
LinearSVC	5.0	0.16-0.60	0.099	0.094	10758.510
Log-Odds	5.0	0.16-0.60	0.105	0.101	8888.560
TF-IDF	5.0	0.16-0.60	0.095	0.090	11987.110
TF-IDF Neg-Pos	5.0	0.16-0.60	0.079	0.076	14931.920
XGBoost	5.0	0.16-0.60	0.102	0.098	10421.210
Cluster	3.0	0.20-0.50	0.102	0.096	6947.640
LinearSVC	3.0	0.20-0.50	0.105	0.099	6585.650
Log-Odds	3.0	0.20-0.50	0.111	0.105	5309.480
TF-IDF	3.0	0.20-0.50	0.100	0.095	6869.440
TF-IDF Neg-Pos	3.0	0.20-0.50	0.081	0.077	8252.790
XGBoost	3.0	0.20-0.50	0.108	0.102	5846.400
None	3.0	0.15-0.15	0.111	0.105	5131.950

Table A.3: DSPT all results for anxiety domain using BERT as base model

Masking Strategy	Epochs	Min-Max MP	Loss Score		Time (s)
			Train	Validation	
Absolutist Words	3.0	0.35-0.35	0.0746	0.070	5631.870
Absolutist Words	3.0	0.50-0.50	0.0744	0.070	5568.010
Absolutist Words	3.0	0.65-0.65	0.0739	0.070	5924.100
Cluster	3.0	0.16-0.60	0.0686	0.064	7775.500
LinearSVC	3.0	0.16-0.60	0.0692	0.065	7440.810
Log-Odds	3.0	0.16-0.60	0.0749	0.070	5791.940
TF-IDF	3.0	0.16-0.60	0.0667	0.062	7863.360
TF-IDF Neg-Pos	3.0	0.16-0.60	0.0566	0.053	10475.110
XGBoost	3.0	0.16-0.60	0.0723	0.067	6772.430
Absolutist Words	5.0	0.35-0.35	0.0713	0.068	9766.750
Absolutist Words	5.0	0.50-0.50	0.0722	0.068	9466.480
Absolutist Words	5.0	0.65-0.65	0.0711	0.068	9856.430
Cluster	5.0	0.16-0.60	0.0663	0.063	13432.450
LinearSVC	5.0	0.16-0.60	0.0664	0.063	12440.870
Log-Odds	5.0	0.16-0.60	0.0714	0.068	10484.100
TF-IDF	5.0	0.16-0.60	0.0654	0.062	13985.320
TF-IDF Neg-Pos	5.0	0.16-0.60	0.0537	0.052	16469.390
XGBoost	5.0	0.16-0.60	0.0697	0.066	11281.640
Cluster	3.0	0.20-0.50	0.0683	0.064	8287.560
LinearSVC	3.0	0.20-0.50	0.0704	0.065	6889.550
Log-Odds	3.0	0.20-0.50	0.0742	0.070	6030.630
TF-IDF	3.0	0.20-0.50	0.0667	0.062	7708.680
TF-IDF Neg-Pos	3.0	0.20-0.50	0.0535	0.050	9678.390
XGBoost	3.0	0.20-0.50	0.0726	0.068	6859.260
None	3.0	0.15-0.15	0.0741	0.070	5949.400

Table A.4: DSPT all results for anxiety domain using RoBERTa as base model

A. APPENDIX

Masking Strategy	Epochs Min-Max MP	Reddit Dataset		Twitter Dataset		
		Train Accuracy	Validation Accuracy	Precision	Recall	F1-Score
Absolutist Words	3.0 0.35-0.35	0.988	0.976	0.789	0.732	0.759
Absolutist Words	3.0 0.50-0.50	0.987	0.973	0.830	0.798	0.814
Absolutist Words	3.0 0.65-0.65	0.987	0.981	0.781	0.684	0.729
Cluster	3.0 0.16-0.60	0.987	0.981	0.780	0.858	0.817
LinearSVC	3.0 0.16-0.60	0.985	0.978	0.843	0.781	0.811
Log-Odds	3.0 0.16-0.60	0.986	0.978	0.837	0.733	0.782
TF-IDF	3.0 0.16-0.60	0.986	0.978	0.803	0.770	0.786
TF-IDF Neg-Pos	3.0 0.16-0.60	0.987	0.973	0.760	0.877	0.815
XGBoost	3.0 0.16-0.60	0.984	0.971	0.841	0.735	0.784
Absolutist Words	5.0 0.35-0.35	0.985	0.970	0.862	0.571	0.687
Absolutist Words	5.0 0.50-0.50	0.987	0.980	0.756	0.876	0.811
Absolutist Words	5.0 0.65-0.65	0.989	0.978	0.818	0.685	0.746
Cluster	5.0 0.16-0.60	0.987	0.980	0.786	0.700	0.741
LinearSVC	5.0 0.16-0.60	0.985	0.977	0.801	0.685	0.739
Log-Odds	5.0 0.16-0.60	0.987	0.968	0.824	0.661	0.733
TF-IDF	5.0 0.16-0.60	0.987	0.971	0.830	0.575	0.679
TF-IDF Neg-Pos	5.0 0.16-0.60	0.989	0.975	0.844	0.591	0.695
XGBoost	5.0 0.16-0.60	0.988	0.965	0.872	0.641	0.739
Cluster	3.0 0.20-0.50	0.987	0.977	0.832	0.704	0.762
LinearSVC	3.0 0.20-0.50	0.989	0.971	0.839	0.752	0.793
Log-Odds	3.0 0.20-0.50	0.988	0.966	0.796	0.603	0.686
TF-IDF	3.0 0.20-0.50	0.987	0.971	0.845	0.651	0.735
TF-IDF Neg-Pos	3.0 0.20-0.50	0.987	0.969	0.772	0.459	0.575
XGBoost	3.0 0.20-0.50	0.986	0.975	0.853	0.680	0.757
None	3.0 0.15-0.15	0.985	0.971	0.832	0.836	0.834

Table A.5: Fine-Tuning all results for depression domain using BERT as a base model

Masking Strategy	Epochs Min-Max MP	Reddit Dataset		Twitter Dataset		
		Train Accuracy	Validation Accuracy	Precision	Recall	F1-Score
Absolutist Words	3.0 0.35-0.35	0.983	0.960	0.831	0.505	0.628
Absolutist Words	3.0 0.50-0.50	0.982	0.978	0.848	0.472	0.606
Absolutist Words	3.0 0.65-0.65	0.981	0.979	0.745	0.803	0.773
Cluster	3.0 0.16-0.60	0.984	0.980	0.769	0.727	0.747
LinearSVC	3.0 0.16-0.60	0.988	0.977	0.798	0.629	0.704
Log-Odds	3.0 0.16-0.60	0.985	0.979	0.824	0.859	0.841
TF-IDF	3.0 0.16-0.60	0.985	0.982	0.779	0.624	0.693
TF-IDF Neg-Pos	3.0 0.16-0.60	0.985	0.965	0.843	0.399	0.542
XGBoost	3.0 0.16-0.60	0.984	0.973	0.856	0.641	0.733
Absolutist Words	5.0 0.35-0.35	0.984	0.982	0.823	0.733	0.776
Absolutist Words	5.0 0.50-0.50	0.983	0.974	0.840	0.546	0.662
Absolutist Words	5.0 0.65-0.65	0.985	0.978	0.778	0.795	0.786
Cluster	5.0 0.16-0.60	0.983	0.978	0.779	0.892	0.832
LinearSVC	5.0 0.16-0.60	0.983	0.953	0.870	0.533	0.661
Log-Odds	5.0 0.16-0.60	0.982	0.975	0.796	0.583	0.673
TF-IDF	5.0 0.16-0.60	0.985	0.981	0.799	0.854	0.826
TF-IDF Neg-Pos	5.0 0.16-0.60	0.872	0.923	0.886	0.257	0.398
XGBoost	5.0 0.16-0.60	0.979	0.980	0.762	0.843	0.800
Cluster	3.0 0.20-0.50	0.983	0.979	0.796	0.762	0.778
LinearSVC	3.0 0.20-0.50	0.983	0.977	0.775	0.863	0.817
Log-Odds	3.0 0.20-0.50	0.984	0.977	0.750	0.571	0.648
TF-IDF	3.0 0.20-0.50	0.980	0.941	0.862	0.238	0.374
TF-IDF Neg-Pos	3.0 0.20-0.50	0.982	0.981	0.770	0.758	0.764
XGBoost	3.0 0.20-0.50	0.984	0.984	0.811	0.798	0.805
None	3.0 0.15-0.15	0.982	0.914	0.857	0.467	0.605

Table A.6: Fine-Tuning all results for depression domain using RoBERTa as a base model

Masking Strategy	Epochs Min-Max MP	Reddit Dataset		Twitter Dataset		
		Train Accuracy	Validation Accuracy	Precision	Recall	F1-Score
Absolutist Words	3.0 0.35-0.35	0.987	0.973	0.884	0.897	0.890
Absolutist Words	3.0 0.50-0.50	0.987	0.941	0.948	0.537	0.685
Absolutist Words	3.0 0.65-0.65	0.985	0.974	0.889	0.974	0.930
Cluster	3.0 0.16-0.60	0.985	0.977	0.893	0.956	0.924
LinearSVC	3.0 0.16-0.60	0.987	0.976	0.893	0.952	0.922
Log-Odds	3.0 0.16-0.60	0.987	0.974	0.951	0.787	0.861
TF-IDF	3.0 0.16-0.60	0.987	0.940	0.942	0.415	0.577
TF-IDF Neg-Pos	3.0 0.16-0.60	0.988	0.964	0.978	0.662	0.789
XGBoost	3.0 0.16-0.60	0.986	0.978	0.946	0.897	0.921
Absolutist Words	5.0 0.35-0.35	0.986	0.975	0.904	0.904	0.904
Absolutist Words	5.0 0.50-0.50	0.988	0.971	0.939	0.912	0.925
Absolutist Words	5.0 0.65-0.65	0.987	0.974	0.872	0.949	0.908
Cluster	5.0 0.16-0.60	0.984	0.973	0.871	0.897	0.884
LinearSVC	5.0 0.16-0.60	0.987	0.973	0.878	0.949	0.912
Log-Odds	5.0 0.16-0.60	0.987	0.977	0.890	0.890	0.890
TF-IDF	5.0 0.16-0.60	0.983	0.972	0.847	0.897	0.871
TF-IDF Neg-Pos	5.0 0.16-0.60	0.987	0.978	0.923	0.930	0.927
XGBoost	5.0 0.16-0.60	0.987	0.977	0.874	0.945	0.908
Cluster	3.0 0.20-0.50	0.987	0.976	0.866	0.930	0.897
LinearSVC	3.0 0.20-0.50	0.985	0.967	0.926	0.732	0.817
Log-Odds	3.0 0.20-0.50	0.987	0.977	0.877	0.868	0.872
TF-IDF	3.0 0.20-0.50	0.987	0.959	0.967	0.640	0.770
TF-IDF Neg-Pos	3.0 0.20-0.50	0.989	0.978	0.895	0.912	0.903
XGBoost	3.0 0.20-0.50	0.986	0.974	0.869	0.879	0.874
None	3.0 0.15-0.15	0.988	0.926	0.962	0.563	0.710

Table A.7: Fine-Tuning all results for anxiety domain using BERT as a base model

Masking Strategy	Epochs Min-Max MP	Reddit Dataset		Twitter Dataset		
		Train Accuracy	Validation Accuracy	Precision	Recall	F1-Score
Absolutist Words	3.0 0.35-0.35	0.985	0.977	0.924	0.846	0.883
Absolutist Words	3.0 0.50-0.50	0.985	0.979	0.893	0.956	0.924
Absolutist Words	3.0 0.65-0.65	0.986	0.978	0.923	0.919	0.921
Cluster	3.0 0.16-0.60	0.987	0.979	0.934	0.886	0.909
LinearSVC	3.0 0.16-0.60	0.987	0.978	0.840	0.949	0.891
Log-Odds	3.0 0.16-0.60	0.983	0.972	0.825	0.691	0.752
TF-IDF	3.0 0.16-0.60	0.985	0.977	0.832	0.945	0.885
TF-IDF Neg-Pos	3.0 0.16-0.60	0.986	0.979	0.885	0.930	0.907
XGBoost	3.0 0.16-0.60	0.983	0.979	0.933	0.875	0.903
Absolutist Words	5.0 0.35-0.35	0.983	0.973	0.821	0.978	0.893
Absolutist Words	5.0 0.50-0.50	0.984	0.973	0.967	0.746	0.842
Absolutist Words	5.0 0.65-0.65	0.986	0.971	0.963	0.757	0.848
Cluster	5.0 0.16-0.60	0.985	0.969	0.810	0.941	0.871
LinearSVC	5.0 0.16-0.60	0.986	0.975	0.793	0.702	0.745
Log-Odds	5.0 0.16-0.60	0.985	0.974	0.915	0.754	0.827
TF-IDF	5.0 0.16-0.60	0.984	0.975	0.894	0.897	0.895
TF-IDF Neg-Pos	5.0 0.16-0.60	0.984	0.972	0.884	0.978	0.928
XGBoost	5.0 0.16-0.60	0.985	0.982	0.894	0.901	0.897
Cluster	3.0 0.20-0.50	0.983	0.977	0.890	0.919	0.904
LinearSVC	3.0 0.20-0.50	0.498	0.508	0.000	0.000	0.000
Log-Odds	3.0 0.20-0.50	0.985	0.978	0.916	0.926	0.921
TF-IDF	3.0 0.20-0.50	0.986	0.978	0.951	0.860	0.903
TF-IDF Neg-Pos	3.0 0.20-0.50	0.982	0.975	0.861	0.982	0.918
XGBoost	3.0 0.20-0.50	0.985	0.947	0.954	0.684	0.797
None	3.0 0.15-0.15	0.984	0.978	0.879	0.879	0.879

Table A.8: Fine-Tuning all results for anxiety domain using RoBERTa as a base model

List of Figures

1.1 Proposed Architecture: Add DSPT between general pre-training and fine-tuning	2
2.1 Seq2Seq structure [21] and Transformers architecture [22]	7
2.2 Scaled Dot-Product Attention and Multi-Head Attention of transformers [22]	8
2.3 BERT architecture: pre-training and fine-tuning [11]	9
3.1 Shannon Entropy Shifts between Reddit data and Wikipedia+Bookcorpus	16
3.2 Shannon Entropy Shifts between Reddit and Twitter data	17
3.3 Pre-processing steps for Reddit and Twitter data	20
4.1 Binary Cross Entropy Loss for Log-Odds strategy	31
4.2 DSPT model structure	32
4.3 Binary Cross Entropy Loss for Log-Odds strategy	36
4.4 Percentage of false positives in different subreddits	39
4.5 Global explanation when using LIME for depression domain	40
4.6 Global explanation when using SHAP for depression domain	41
4.7 Global explanation when using LIME for anxiety domain	42
4.8 Global explanation when using SHAP for anxiety domain	42
4.9 Local explanation when using LIME for depression domain	43
4.10 Local explanation when using SHAP for depression domain	44
4.11 Local explanation when using LIME for anxiety domain	45
4.12 Local explanation when using SHAP for anxiety domain	46
A.1 First 2 replies from Reddit admins regarding Reddit data usage	52
A.1 Last 2 replies from Reddit admins regarding Reddit data usage	53
A.2 Dictionary-based sentiment analysis for Reddit and Wikipedia+Bookcorpus for depression domain	54
A.3 Dictionary-based sentiment analysis for Reddit and Twitter data for depression domain	55
A.4 Shannon Entropy Shifts between Twitter and Wikipedia+Bookcorpus for depression domain	56
A.5 Dictionary-based sentiment analysis for Twitter and Wikipedia+Bookcorpus for depression domain	57
	69

A.6 Shannon Entropy Shifts between Reddit and Wikipedia+Bookcorpus for anxiety domain	58
A.7 Dictionary-based sentiment analysis for Reddit and Wikipedia+Bookcorpus for anxiety domain	59
A.8 Shannon Entropy Shifts between Reddit and Twitter data for anxiety domain	60
A.9 Dictionary-based sentiment analysis for Reddit and Twitter data for anxiety domain	61
A.10 Shannon Entropy Shifts between Twitter and Wikipedia+Bookcorpus for anxiety domain	62
A.11 Dictionary-based sentiment analysis for Twitter and Wikipedia+Bookcorpus for anxiety domain	63

List of Tables

3.1	Number of posts and average number of words per post for each subreddit data	14
3.2	Number of tweets and average number of words per tweet	16
4.1	DSPT results for depression domain using BERT as base model	29
4.2	Number of words which were present in tokenizer for each masking strategy for depression domain	30
4.3	DSPT results for depression domain using RoBERTa as base model	30
4.4	Fine-Tuning best results for depression domain using BERT as a base model for all masking strategies	32
4.5	Fine-Tuning best results for depression domain using RoBERTa as a base model for all masking strategies	33
4.6	Average scores and standard deviation for each Masking Strategy over all trained models for depression domain	34
4.7	DSPT results for anxiety domain using BERT as base model	35
4.8	Nr. of words which were present in tokenizer for each masking strategy for anxiety domain	35
4.9	DSPT results for anxiety domain using RoBERTa as base model	36
4.10	Fine-Tuning best results for anxiety domain using BERT as a base model for all masking strategies	37
4.11	Fine-Tuning best results for anxiety domain using RoBERTa as a base model for all masking strategies	38
4.12	Average scores and standard deviation for each Masking Strategy over all trained models for anxiety domain	38
5.1	Recall scores for baseline models and best stable models	49
A.1	DSPT all results for depression domain using BERT as base model	64
A.2	DSPT all results for depression domain using RoBERTa as base model	64
A.3	DSPT all results for anxiety domain using BERT as base model	65
A.4	DSPT all results for anxiety domain using RoBERTa as base model	65
A.5	Fine-Tuning all results for depression domain using BERT as a base model	66
A.6	Fine-Tuning all results for depression domain using RoBERTa as a base model	66
A.7	Fine-Tuning all results for anxiety domain using BERT as a base model	67
		71

List of Algorithms

1	Mask Words	24
---	----------------------	----

Bibliography

- [1] WHO. Depression and other common mental disorders: Global health estimates. <https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf>, 2017. [Accessed 22-Apr-2023].
- [2] WHO. Mental health and covid-19: Early evidence of the pandemic’s impact. https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci_Brief-Mental_health-2022.1, 2022. [Accessed 22-Apr-2023].
- [3] Lara B Akin, Jan-Emmanuel De Neve, Elizabeth W Dunn, Daisy E Fancourt, Elkhonon Goldberg, John F Helliwell, Sarah P Jones, Elie Karam, Richard Layard, Sonja Lyubomirsky, Andrew Rzepa, Shekhar Saxena, Emily M Thornton, Tyler J VanderWeele, Ashley V Whillans, Jamil Zaki, Ozge Karadag, and Yanis Ben Amor. Mental health during the first year of the COVID-19 pandemic: A review and recommendations for moving forward. *Perspect. Psychol. Sci.*, 17(4):915–936, July 2022.
- [4] Donatella Marazziti, Maria T Avella, Nicola Mucci, Alessandra Della Vecchia, Tea Ivaldi, Stefania Palermo, and Federico Mucci. Impact of economic crisis on mental health: a 10-year challenge. *CNS Spectr.*, 26(1):7–13, February 2021.
- [5] Harriet E Ingle and Michael Mikulewicz. Mental health and climate change: tackling invisible injustice. *Lancet Planet. Health*, 4(4):e128–e130, April 2020.
- [6] Stevie Chancellor and Munmun De Choudhury. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ Digit. Med.*, 3(1):43, March 2020.
- [7] Tiberiu Sosea and Cornelia Caragea. eMLM: A new pre-training objective for emotion related tasks. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 286–293, Online, August 2021. Association for Computational Linguistics.
- [8] Nikolay Arefyev, Dmitrii Kharchev, and Artem Shelmanov. Nb-mlm: Efficient domain adaptation of masked language models for sentiment analysis. *Proceedings*

of the 2021 Conference on Empirical Methods in Natural Language Processing, page 914, November 2021.

- [9] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. Mentalbert: Publicly available pretrained language models for mental healthcare. *CoRR*, abs/2110.15621, 2021.
- [10] Adrian Zbiciak and Tymon Markiewicz. A new extraordinary means of appeal in the polish criminal procedure: the basic principles of a fair trial and a complaint against a cassatory judgment. *Access to Justice in Eastern Europe*, 6(2):1–18, March 2023.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.
- [12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.
- [13] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado, June 5 2015. Association for Computational Linguistics.
- [14] Sklearn. LinearSVC. <https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html>. [Accessed 10-Nov-2022].
- [15] XGBoost. Xgboost. <https://xgboost.readthedocs.io/en/stable/>. [Accessed 10-Nov-2022].
- [16] Huggingface. Tokenizer. https://huggingface.co/docs/transformers/main_classes/tokenizer. [Accessed 17-Nov-2022].
- [17] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909, 2015.
- [18] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016.
- [19] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017.
- [20] Koustuv Saha, Benjamin Sugar, John Torous, Bruno Abrahao, Emre Kiciman, and Munmun De Choudhury. A social media study on the effects of psychiatric medication use. *Proceedings of the International AAAI Conference on Web and Social Media*, 13:440–451, July 2019.

- [21] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014.
- [22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.
- [23] Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *CoRR*, abs/1506.06724, 2015.
- [24] Sebastian Nagel. Cc news. <https://commoncrawl.org/2016/10/news-dataset-available/>, 2016. [Accessed 21-apr-2023].
- [25] Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. <http://Skylion007.github.io/OpenWebTextCorpus>, 2019.
- [26] Trieu H. Trinh and Quoc V. Le. A simple method for commonsense reasoning. *CoRR*, abs/1806.02847, 2018.
- [27] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. *CoRR*, abs/2004.10964, 2020.
- [28] Yuxian Gu, Zhengyan Zhang, Xiaozhi Wang, Zhiyuan Liu, and Maosong Sun. Train no evil: Selective masking for task-guided pre-training. *CoRR*, abs/2004.09733, 2020.
- [29] Nikolay Arefyev, Dmitrii Kharchev, and Artem Shelmanov. NB-MLM: Efficient domain adaptation of masked language models for sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9114–9124, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [30] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. Mentalbert: Publicly available pretrained language models for mental healthcare. *CoRR*, abs/2110.15621, 2021.
- [31] Reddit. Api. <https://www.reddit.com/dev/api/>. [Accessed 24-apr-2023].
- [32] Ryan J Gallagher, Morgan R Frank, Lewis Mitchell, Aaron J Schwartz, Andrew J Reagan, Christopher M Danforth, and Peter Sheridan Dodds. Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts. *EPJ Data Sci.*, 10(1), December 2021.

- [33] Mohammed Al-Mosaiwi and Tom Johnstone. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, 6(4):529–542, 2018. PMID: 30886766.
- [34] Twitter. Developer policy. <https://developer.twitter.com/en/developer-terms/policy>. [Accessed 25-apr-2023].
- [35] Twitter. More about restricted uses of the twitter apis. <https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases>. [Accessed 25-apr-2023].
- [36] Reddit. Developer terms. <https://www.redditinc.com/policies/developer-terms>. [Accessed 25-apr-2023].
- [37] Dredze Mark. Clpsych 2015 - data use and confidentiality agreement. <https://www.cs.jhu.edu/~mdredze/clpsych-2015-shared-task-evaluation/>. [Accessed 23-apr-2023].
- [38] GDPR. Article 89. https://gdprhub.eu/Article_89_GDPR. [Accessed 04-mar-2023].
- [39] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909, 2015.
- [40] Enja Kokalj, Blaž Škrlić, Nada Lavrač, Senja Pollak, and Marko Robnik-Šikonja. BERT meets shapley: Extending SHAP explanations to transformer-based classifiers. In Hannu Toivonen and Michele Boggia, editors, *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 16–21, Online, April 2021. Association for Computational Linguistics.