Tracing Subtle Patterns: Early Detection of Anorexia in Social Media Using Contextualized BERT Embeddings and Deep Learning

by

Quazi Fariha Fairooz
24141259
Tawhidur Rahman Khan
24141239

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
October 2024

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Quazi Fariha Fairooz	Tawhidur Rahman Khan
24141259	24141239

Approval

The thesis/project titled "Tracing Subtle Patterns: Early Detection of Anorexia in Social Media Using Contextualized BERT Embeddings and Deep Learning" submitted by

1. Quazi Fariha Fairooz (24141259)
2. Tawhidur Rahman Khan (24141239)

Of Summer, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in October, 2024.

Examining Committee:

Supervisor:
(Member)

Dr. Farig Yousuf Sadeque (PhD)
Associate Professor
Department of Computer Science and Engineering
BRAC University

Co-Supervisor:
(Member)

Jawaril Munshad Abedin
Lecturer
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam (PhD)
Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi (PhD)
Chairperson and Associate Professor
Department of Computer Science and Engineering
BRAC University

Abstract

In a world where the usage of social media is prevalent, it is important to acknowledge the fact that these social platforms have been increasingly used to share private matters with the public eye, which may sometimes reveal underlying symptoms of diseases unknown to the user. Mental health disorder is a severe issue and with the growing popularity of these social media platforms, online bullying has also been on the rise, leading to users being stressed mentally and developing various mental disorders. In this paper, we tried to evaluate four state-of-the-art deep learning models: LSTM, BiLSTM, GRU, and BiGRU, along with BERT, to detect eating disorder-related content, specifically anorexia, in online communities. Using a large-scale dataset extracted from Reddit, our research mainly focused on the early detection of anorexia from the posts/comments of users. We aimed to use the aforementioned deep learning models in such a way that would surpass the performance of other models that were used on our dataset and also, result in a significant advancement in the field of NLP. Our approach involved an extensive analysis of linguistic cues, semantic context, and user-generated content to identify both subtle and explicit mentions of anorexia in social media texts. We tried to attain a higher positive recall and detect anorexia faster than the models used on the dataset by incorporating the latest techniques and combining various elements from existing approaches with the best performance in detecting indicators of this potentially life-threatening eating disorder. With our models, we were able to achieve a high positive class recall of 0.93, and our best values of $ERDE_5$ and $ERDE_{50}$ were 12.63% and 6.99% respectively.

Keywords: Deep Learning; Machine Learning; NLP; BERT; Eating Disorder; Anorexia, Early Detection

Table of Contents

Declaration **i**

Approval **ii**

Abstract **iv**

Table of Contents **v**

List of Figures **vii**

List of Tables **1**

1 **2**
1.1 Introduction 2
1.2 Problem Statement 2
1.3 Research Objectives 3

2 Related Work **4**

3 Dataset **13**
3.1 Analysis of Dataset 13
3.2 Data Preprocessing 15

4 Model **19**
4.1 Long Short-Term Memory (LSTM) 19
4.1.1 Bi-Directional Long Short-Term Memory (Bi-LSTM) 20
4.2 Gated Recurrent Unit (GRU) 21
4.2.1 Bi-Directional Gated Recurrent Unit (BiGRU) 21
4.3 BERT 23
4.3.1 BERT's Mechanism 23
4.3.2 BERT Encoder: Input Dimensions and Processing 24

5 Experiment and Result Analysis **28**
5.1 LSTM Training Evaluation 30
5.2 BiLSTM Training Evaluation 30
5.3 GRU Training Evaluation 32
5.4 BiGRU Training Evaluation 32
5.5 Combined Testing Evaluation and Analysis 33
5.5.1 Evaluation Metrics 33
5.6 Performance Comparison With Previous Results 34

6 Limitations and Challenges **36**

7 Conclusion **38**
7.1 Future Work 38

Bibliography **41**

List of Figures

3.1 Row Distribution 13
3.2 CLEF eRisk 2018 Anorexia Dataset Statistics 14
3.3 First 10 Rows of Combined Dataframe After Changing The Column Names 14
3.4 TEXT Column's Row Length Distribution 15
3.5 Top 10 Words In Anorexia-Positive and Anorexia-Negative Datasets . 16
3.6 Word Clouds for Anorexia-Positive and Anorexia-Negative Datasets . 17
3.7 Word Cloud For Combined Dataset After Preprocessing 17

4.1 A single LSTM Recurrent Unit [23] 19
4.2 Bi-LSTM Architecture [1] 20
4.3 A Single GRU Unit [11] 21
4.4 BiGRU Architecture [12] 22
4.5 How BERT Works [16] 24
4.6 BERT Base and BERT Large Encoder Architecture [25] 26

5.1 Using BERT To Generate Contextualized Post Embeddings and Feed It To LSTM 28
5.2 Detection Threshold Mechanism 30
5.3 Training and Validation Loss of LSTM Over Epochs 31
5.4 Training and Validation Loss of BiLstm Over Epochs 31
5.5 Training and Validation Loss of GRU Over Epochs 32
5.6 Training and Validation Loss of BiGRU Over Epochs 33

List of Tables

5.1 Performance Comparison of Different Models 34

Chapter 1

1.1 Introduction

Eating disorders are behavioral and mental health conditions defined by profound, prolonged, and persistent disturbances in eating habits. The DSM-5 (Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition) categorizes eating disorders as "Feeding & Eating Disorders" and defines them as "characterized by a persistent disturbance of eating or eating-related behavior that results in altered consumption or absorption of food that significantly impairs physical health or psychosocial functioning" [28].

There are multiple types of eating disorders like Anorexia Nervosa, which indicates a significant reduction in food intake; Binge eating disorders (BED), mostly characterized by overeating or emotional eating; Eating disorders not otherwise specified (EDNOS); Night-eating syndrome (NES); Bulimia Nervosa, and many others. In our research, we will be dealing with Anorexia Nervosa.

Anorexia Nervosa is a potentially life-threatening but curable, if detected early, eating disorder. Sufferers of anorexia extremely limit their calorie count because of the fear of gaining weight which eventually results in malnutrition. Study shows that patients with anorexia have a risk of suicide 18 times higher than others[26].

1.2 Problem Statement

Eliminating the problem of insufficient datasets when identifying anorexia-related texts, the study aims to propose a comparison analysis of the integration of BERT with a variety of deep learning models and find the best combination for the early detection of anorexia with increased efficiency.

To elaborate, a statistic provided by ANAD (the National Association of Anorexia Nervosa and Associated Disorders) shows that 22% of children and adolescents worldwide indicate eating behaviors that are unhealthy. Patients with eating disorders have a risk of suicide 11 times higher than those without an eating disorder [26]. Due to these eating disorders, people suffer from side effects like osteoporosis because of lowered intake of food and therefore, decreased estrogen related to amenorrhea, damage of taste receptors due to purging which is an effect of bulimia nervosa, severe

weight change, fatigue, and others. Thus, detecting this fatal disease as early as possible is a necessity. However, the detection of eating disorders in online discourse is a challenging task since the content posted by users online contains only subtle hints of their emotions and although there are existing models that manage to complete the classification task, they exhibit limitations in identifying anorexia-related posts, which calls for the necessity of better approaches to be developed quickly - a goal we tried to achieve in our research. Additionally, not all classification tasks have sufficient annotated datasets available on the Internet, which points out a critical gap in NLP technologies, prompting the need for an approach to solve the problem of insufficient datasets, which we also tried to accomplish in this study.

1.3 Research Objectives

In the case of deep learning-based approaches, our future goals include not just exploring new models to increase their efficiency on a task, but also explaining the behavior of these models. Moreover, using these models we would like to achieve and demonstrate the subtle patterns of eating disorders. We aim to represent how text data can be used as an indicator of a potential disorder. We anticipate achieving more accurate text classification methods and comparing the effectiveness of text classification techniques by utilizing RNN and transformer-based models. Furthermore, we would like to determine the optimal technique for future research by balancing performance and computational costs. We aspire to demonstrate that as due to the emergence of social media platforms, identifying and finding the patterns of diseases with the help of text classification methods has become quite convenient. For future work, we aspire to evaluate the models in an experimental context to inform treatment selections, incorporate Natural Language Processing techniques to utilize ontologies and enable automation and logical reasoning, and integrate predictive models into practical development projects while validating them. Lastly, we also want to reduce errors produced by the models implemented in eating disorder-specific classification tasks.

To sum up, our research objectives are:

- Use deep learning models to detect anorexia in social media posts
- Maximize the positive recall during the testing phase
- Surpass the results produced by other techniques used on the same dataset
- Ensure early and accurate detection of anorexia for quicker and more reliable warnings
- Get our paper published in a renowned journal to help guide people interested in this field and broaden the scope of our work.

Chapter 2

Related Work

Researchers have explored and provided several approaches to detect eating disorders. Some of those have performed tremendously well and after that scientists not only studied those techniques but also improved the existing models throughout time with the help of acquired knowledge. To obtain comprehensive knowledge about the existing models and algorithms, we studied some of the most cited research papers from Google Scholar. We believe this knowledge will help us to propose an approach that will enhance future opportunities in the field of data science.

To begin with, the authors (Wang et al.) of [5] proposed an approach that used a combination of TF-IDF (Term Frequency-Inverse Document Frequency) information and convolutional neural networks (CNNs) to detect depression and anorexia from social media texts. The datasets of anorexia and depression were collected from eRisk 2018 which consisted of posts and comments on Reddit. The authors (Wang et al.) considered the problem as a classification task and to solve it, they first calculated the TF-IDF of each word in both datasets and removed the words with less TF-IDF. After that, CNN classifier was trained on the datasets. To predict the results, the authors (Wang et al.) used thresholds $\theta 1$, $\theta 2$, and $\theta 3$, which ranged between 0 and 1. They identified a person as risky if above $\theta 1$ of their terms was classified as positive, i.e. as depressed or anorexic. On the contrary, if less than $\theta 2$ posts/comments were tagged as negative (not depressed or non-anorexic), they considered the user to be non-risky. Otherwise, the last portion of the collection of the user's posts/comments was used to label them - if above $\theta 3$ of terms was tagged as positive, the user was identified as risky, or else the user was given a non-risky tag. Furthermore, in depression identification, Early Risk Detection Error (ERDE), in this case, ERDE_5, was 10.81%, ERDE_{50} was 9.22%, and F-score was 0.37, whereas, in anorexia detection, ERDE_5 was 13.65%, ERDE_{50} was 11.14%, and F-score was 0.67. While comparing their performance with other leading teams for the CLEF eRisk 2018 task, it was shown that this combination of TF-IDF and CNNs based model performed better in anorexia detection based on ERDE_5.

Sometimes researchers also build their corpus or dataset and make it available for the benefit of others. In the paper [6], the authors (Ubeda et al.) presented a new corpus in Spanish for detecting anorexia in Spanish Tweets, and they also compared the performance of the machine learning algorithms: Naive Bayes (NB), Support

Vector Machine (SVM), Multilayer Perceptron (MLP), Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR), on their own compiled corpus, SAD (Spanish Anorexia Dataset). The corpus consisted of 2,707 Anorexia related tweets and 3,000 NON-Anorexia related tweets, which were achieved by filtering data (removing repeated tweets, all hashtags, and short tweets). These tweets were collected using hashtags and API (Application Programming Interface), which allows users to retrieve tweets in a specific language, in this case, Spanish. For grammatical labeling, the authors (Ubeda et al.) used spaCy, which is a free open-source library for NLP in Python. Pre-Processing was done to enhance the performance of the classifier and speed up the process of classification. For this, NLTK TweetTokenizer was used to tokenize and remove the hashtags, and in order to show importance of each feature, TF-IDF was used to add weight to the texts. For comparing the performance of machine learning algorithms, the authors (Ubeda et al.) of [6] used precision, recall, F-score (F1) and accuracy. A 10-fold cross validation technique was used to evaluate the machine learning classifiers and the best performing classifiers were SVM and MLP obtaining accuracy above 0.9. The other classifiers also presented well enough results, all achieving an accuracy score of up to 80% on a well balanced corpus. Only 8.38% of the total tweets were not correctly classified. The system in [6] recognized terms associated with food, and it also noted that the vocabulary of the tweets classified as control (non-anorexia) was remarkably similar to that of anorexia, which may have resulted in it being occasionally incorrect.

Besides, some existing models are modified by researchers for better performance. Due to the problem of the bag-of-words (BOW) model used by the SS3 classifier where the word sequences were not recognized, the authors (Burdisso et al.) introduced a modified version of the SS3 model in the paper [7], the τ-SS3, which was able to recognize different length word n-grams dynamically and hence, allowed the model to recognize useful patterns in texts, which the older model could not do. The datasets that were used to experiment with the new τ-SS3 model in [7] were the collections of writings that were submitted by Reddit users, which were used for the CLEF's eRisk tasks of eRisk 2017 early depression detection, eRisk 2018 early depression detection and early anorexia detection. As mentioned beforehand, the problem with the SS3 classifier was that it used a simple BOW model, not allowing it to detect patterns and word sequences of concern from the sentences it processed, which in turn adversely affected its classification ability. Another problem was that the simple BOW recognized unigrams, which contain less information than word sequences. The authors (Burdisso et al.) of [7] addressed these drawbacks of the SS3 model by improving its features, enabling the model to identify useful and important word sequences by using a prefix tree to save n-grams of different lengths that were detected in the training phase. Next, the prefix tree was used as a Deterministic Finite Automata (DFA) to detect useful word sequences when the input was looked through. After the τ-SS3 model was trained and tested, it was seen that it performed somewhat better than the older model in all of the above mentioned eRisk tasks. In the tasks, many models were used, among which τ-SS3 achieved the best ERDE_{50} in both the depression detection tasks, although it came second in the anorexia detection task. τ-SS3 also outperformed the other models with its fine ERDE_5 score in both the 2017 anorexia and depression detection tasks. Considering standard measures (F1, precision and recall), the τ-SS3 model additionally

improved the SS3 model's classification achievements. Learning word n-grams from the input sentences helped the τ-SS3 model to perform a bit better than the previous SS3 model in all the three tasks. Furthermore, learned n-grams also contributed to improving visual explanations given by SS3, which were previously less informative.

Although traditional machine learning algorithms worked efficiently in classification tasks, researchers explored new frameworks for this purpose as well. The authors (Funk et al.) of the paper [8] proposed an NLP based framework to predict binge eating behavior using data from Digital Health Interventions (DHI). DHI data were mostly collected from two sources: (1) text data throughout the intervention, such as diary entries and posts, and (2) open correspondence via a messaging system between users and qualified professionals assisting them. 37,228 text snippets and 4285 messages by 372 users were used to explore the performance of the framework in [8]. In order to predict the eating disorder, the authors (Funk et al.) performed two steps of tasks and they implemented these tasks as an R package called Digital Health Interventions Text Analytics (DHITA). The first step was feature engineering, which was a combination of metadata, word usage, word embeddings, Parts of speech tagging, topic models (document's semantic structure), and sentiment analysis. The second step was predictive and inference modeling for which three models were introduced. The level of symptoms over time is mostly explained by the model A by making a connection between neighboring text part and syndrome level. The therapeutic effect was determined by model B using the mean sentiment score of each user. Lastly, Model C inferred the characteristics and the importance of each text snippet. For Model A, AUC of the RF method is 0.72. The RF tended to overfit for across-user learning, while the LR produced superior results (AUC=0.57). As model A worked less on the testing data, this method is not chosen for early treatment. To predict the severity of the disease Model B was used by the authors (Funk et al.). For this, they aggregated the text features using the LASSO regression and mean square error (MSE) method which led to 220 aggregated features for each user. In this study [8], Model C was not used.

The reason we wanted to detect eating disorders as early as possible was to ensure that enough assistance could be offered to the sufferers in little to no time. In the paper [9], the authors (Hwang et al.) aimed to design a "Personalized Support System" (PSS) by examining the behavioral patterns of individuals, who use online socializing networks, with symptoms of emotional eating (EE) disorder so that sufficient support could be provided to these users. The dataset in [9] was built using user-generated textual content from a sub=community of Reddit - /r/loseit. The dataset consisted of 185,950 posts and 3,528,107 comments from July 2014 to May 2018. An expert group comprising a nutritionist with in-depth knowledge of EE behavior and a medical doctor, whose expertise lied in EE diagnosis, was used to label the data as EE, or others. To classify the EE posts, the authors (Hwang et al.) used 5 ML models - NB, DT, SVM, k-nearest neighbor (KNN) algorithm, and Stochastic Gradient Descent (SGD). Next, topic modeling was done where Latent Dirichlet Allocation (LDA) was used to analyze EE behavioral patterns and feedback topics. Accuracy, precision, recall and F1 score were used as the evaluation metrics to determine the best ML classifier. SGD proved to be the best performing

model with an accuracy of 90%, precision of 0.92, and recall and F1 score of 0.91. During the topic modeling process, LDA categorized the EE posts into 4 categories - addressing feelings, sharing physical changes, sharing and asking for dietary information, and sharing dietary strategies; and feedback topics into 5 main types: dietary information, compliments, consolation, Reddit bot, and health information. The study [9] showed the potential of the LDA topic modeling in the medical field as an investigative research tool to detect and classify abnormal behaviors, and the promising opportunities to carry out PSS to provide support for emotional eaters. Furthermore, the study also led to the finding of SGD being able to promptly detect EE posts with a high accuracy.

In addition, acknowledging and addressing the pattern of this disease was equally important. For this, the authors (Uban et al.) of the paper [13] incorporated a hierarchical attention network (HAN), LSTM layers encoded deep learning method and attention of two layers. They (Uban et al.) used the eRisk (2019) Reddit dataset on anorexia. Out of the 1,287 users who posted 823,754 posts, 10.4% were anorexic. The deep learning method, HAN, was based on the post-level encoder and the user-level encoder, which were modeled as LSTMs. For embedding, GloVe was used to capture content features. For observing style features, the BOW method was used to ensure the vector representation of words. LIWC and NRC lexicons were used for emotion and syntactic features. In order to train the model for prediction, all of these features were concatenated and passed to LSTM. Attention weight analysis method was used in [13] to calculate the importance of inputs. Then, to reduce and visualize the output in two dimensions, the authors used the principal component analysis (PCA), and a kernel density estimate (KDE) plot for observation of distribution score. The result in [13] showed three distinguishing clusters: ANO1, ANO2 and control (Non-Anorexia). The authors (Uban et al.) observed that cluster ANO2 was a mixture of anorexia (ANO1) and control, and they considered it as a pattern. Applying the 'Five Factor Model' - openness to experience, agreeableness, conscientiousness, neuroticism, extraversion, the test revealed significant differences (p-value $<.005$) in three characteristics between anorexics and control cases: agreeableness, extraversion, and neuroticism. According to statistical differences between the three clusters, neuroticism (emotional instability) indicated that ANO1 users were experiencing a more serious form of this psychological illness than ANO2 users. Moreover, LIWC and emotion lexicons identified a certain degree of social isolation within cluster ANO1, which was higher than ANO2 and control. As performance metrics, the authors (Uban et al.) of [13] computed the precision, recall, F1-score, and AUROC. Comparing the results of HAN model with baselines such as an LR model with bag-of-word features, and transformer-based models including RoBERTa and AlBERT with word sequences as features, the paper [13] proved that HAN model achieved the best results in terms of AUROC.

Remarkably, most studies suggested that machine learning algorithms worked better for identifying eating disorders so, starting in 2021, researchers in the data science field applied various algorithms and compared their performance. For prediction-related tasks, the authors (Benitez-Andrades et al.) of the paper [10] developed six models based on the Bidirectional Encoder Representations from Transformers (BERT) models and compared which model scored the best in terms of classify-

ing tweets into two categories - eating disorder related tweets and tweets that were written by people without any eating disorders. The authors (Benitez-Andrades et al.) used web scraping methods to collect a total of 1,085,957 tweets from Twitter between October 20, 2020, and December 26, 2020, out of which 2,000 tweets were manually labeled by an eating disorder specialist into the two categories as mentioned above, and this was the main dataset that was used for the experiments conducted. In the paper [10], the tweets were extracted using T-Hoarder from Twitter. Following this, preprocessing techniques were applied to the initial 1,085,957 tweets obtained, out of which 494,025 tweets were found to be valid and from these, a subset of the 2,000 tweets was built. The authors (Benitez-Andrades et al.) used the scikit-learn library to divide the data into two sets: training and validation. During the training phase, a few arguments were passed where the input data was set to be reprocessed, half-precision floating point format (FP16) was disabled and the number of epochs for the training process was set to four. Using these arguments six BERT models, namely BERT, RoBERTa, DistilBERT, CamemBERT, ALBERT, and FlauBERT were used for the classification task, with all the models being pre-trained. The model that came out on top with the highest accuracy (87.5%) was the pre-trained RoBERTa model, followed by the pre-trained DistilBERT and CamemBERT models, which had an accuracy of 86.3% and 85.6% respectively. Although the final dataset of 2,000 tweets used was quite small for training deep learning architectures, the performance of the six BERT models that were implemented turned out to be rather satisfactory, which indicates the potential of such predictive models that can be further improved in the future for not only the detection of eating disorders but for predicting other forms of mental illnesses and different classes outside the realm of mental disorders as well.

In another study [14], the authors (Benitez-Andrades et al.) compared traditional machine learning models with BERT models to determine the best approach, combining the performance and computational cost, on four different text classification tasks. From the dataset available on the Kaggle platform, 1,085,957 tweets and after preprocessing, a total of 494,025 valid tweets, from which a subset of 2,000 manually labeled tweets were categorized into four different categories. These categories could be explained as, suffering from eating disorders or not, promoting eating disorders or not, informative or non-informative, and lastly, scientific or not scientific. For the comparison purpose, bidirectional LSTM networks, recurrent neural networks (RNNs), RF, and pre-trained BERT models (RoBERTa, BERT, CamemBERT, DistilBERT, FlauBERT, ALBERT, and RobBERT) were used by the authors (Benitez-Andrades et al.) of [14]. For identifying the people who suffered from eating disorders, RoBERTa and RNN performed the best with an accuracy of 83.1% and 82.6% respectively. Furthermore, achieving the accuracy score of 88.5% and 86.7% respectively, RoBERTa and bidirectional LSTM proved to be the best to identify if the tweet was promoting eating disorders or not. RoBERTa was the best accurate model for detecting informative tweets (accuracy 84.4%). Except for ALBERT and FlauBERT, the accuracy of all BERT models were above 80%, nevertheless, using bidirectional LSTM 78.7% accuracy was achieved. The RoBERTa model (accuracy 94.2%) had the greatest accuracy for detecting scientific or non-scientific tweets. So, the accuracy of all BERT models was 92% or higher; however, using bidirectional LSTM provided an accuracy of 85.8%. The paper [14] proved

that despite the computational expense of BERT models being much greater than other machine learning approaches, they performed significantly better.

Subsequently, transfer learning techniques are also a machine learning approach. The concept is that the information obtained from completing the first task may be applied to learning the second task, potentially leading to greater performance and faster convergence. In the paper [17], the authors (Uban et al.) used transfer learning with linguistic features and multi-aspect representations of languages to detect the four mental disorders: anorexia, depression, PTSD, and self-harm from social media platforms using several deep learning models. The authors (Uban et al.) tried to prove that for disorders that have very limited annotated datasets available, transfer learning could help enhance the prediction performance in such cases. The paper [17] went through various transfer learning strategies for cross-disorder and cross-platform transfer. The authors (Uban et al.) provided feedback on the linguistic features, i.e. which were the most useful ones in transferring knowledge, and this was done by evaluating ablation experiments and by analyzing errors. The datasets used in [17] were: CLEF's eRisk datasets for the tasks of the early detection of anorexia and self-harm (2019 and 2020), and depression (2018), which were a long history of Reddit posts by users; Computational Linguistics and Clinical Psychology (CLPsych) dataset for the task of detecting depression and PTSD (2015), which was a Twitter dataset consisting of users' most recent public Tweets; Twitter dataset on depression (2017), which consisted of Tweets from users that were posted within a month of their self-diagnosis on the disorder. The authors (Uban et al.) recommended and used a complex deep architecture that had several linguistic features, and was based on HAN. The model was implemented in a classification task that comprised multiple classes and multiple labels. Since the datasets were designed for binary classification, the healthy users were not taken into account and the disorders from the two sources, i.e. Reddit and Twitter, were separately classified. Further experiments were conducted using the same HAN model, but using different transfer learning strategies to comprehend the similarities between different disorders (at the language level) and different datasets belonging to distinct social media platforms, and also to improve the accuracy of detecting the disorders individually. From the cross-disorder classification experiments and the variety of transfer learning experiments, it was seen that some disorders were more related to one another than others, and general attributes could be found between the different disorders. Using cross-platform transfer experiments, the authors (Uban et al.) proved that transferring knowledge from datasets with more text to ones with less text could bring forth performance enhancements, while doing the opposite did not bring about any significant improvement.

Moreover, not all datasets contain data written in English. One example was the dataset used in [6], and another example is the paper [15], where the authors (Hacohen-Kerner et al.) aimed to detect anorexic girls from content written in Hebrew that was posted in social media platforms. Five machine learning (ML) models: Support Vector Classifier (SVC), RF, MLP, LR, Multinomial Naive Bayes (MNB); three RNN models; four BERT models: AlephBERT, HeBERT, WikiBERT, BERT multilingual base (cased); three feature filtering techniques (Chi^2, ANOVA, and Mutual Information); three preprocessing rules; parameter tuning; and unique

feature sets (content-based and style-based) were used for different text classification (TC) methods. The authors (Hacohen-Kerner et al.) collected probable anorexia-related blog posts by girls written in Hebrew from blog forums for anorexic girls, and other posts from forums unrelated to any mental disorder. They (Hacohen-Kerner et al.) constructed a dataset from these posts, consisting of 100 anorexia-related posts and 100 non-anorexia-related posts. More importantly, an international expert on anorexia supervised and validated the construction of this dataset. In [15], the authors (Hacohen-Kerner et al.) defined 28 feature sets and they were used on all five of the ML methods as mentioned above. In order to define the finest feature combination for a TC method, all $2^{28}/2$ (134,217,7280) unique combinations of the feature sets had to be taken into account, which was extremely huge. The authors (Hacohen-Kerner et al.) tackled this problem, where n was not a small number, using one of the variants of the hill-climbing process, which had an improved complexity of $O(n^2)$ instead of the problematic $O(2n)$. After obtaining the best combinations of feature sets, i.e. the ones which had the highest values of accuracy, feature filtering and parameter tuning were implemented to further improve the results. The authors (Hacohen-Kerner et al.) of [15] also made one fascinating discovery - traditional ML methods like SVC and RF outperformed deep learning (DL) methods, both in previous challenges like the CLEF eRisk tasks, and also in the experiments that they conducted, and there could be several reasons for this: traditional ML techniques were implemented and went through much more than the DL ones for a longer period and on different datasets; the training phase of DL methods required a comparatively larger dataset size; DL methods were very costly in terms of computational executions and hence, high-end hardware was required; and finally, in order to optimize DL methods, many experiments needed to be carried out using different parameters each time.

More importantly, characterizing the patterns of this condition is equally necessary to correctly recognize the eating disorder. Thus, for the early detection of eating disorders in individuals, the identification of the characteristics of these disorders, and the calculation of the probability of the individuals having the disorders from textual data, the authors (Rujas et al.) of the paper [21] aimed to create an NLP-based method which consisted of three main stages: preprocessing the data using a pipeline, using BERTopic to generate topics, and using a fine-tuned BERT model to develop binary classification and linear regression models. The authors (Rujas et al.) used the dataset that was provided for task 1 of MentalRiskES2023. The dataset comprised a cluster of textual messages sent to groups on the cloud-based messaging application Telegram, where all the messages were labeled, with most of them being written in the Spanish Language. The messages were connected to various eating behaviors, different perspectives on body image, and mental welfare. The authors (Rujas et al.) of [21] used a preprocessing pipeline, consisting of six distinct steps, on the dataset and it was specifically modified to improve the quality of the data and to make the data more presentable and fit for analysis. A topic modeling technique, in this case, the BERTopic method, was used to categorize and cluster eating disorder-related talks by unveiling topics within the textual data that were not directly mentioned. To develop the binary classification and LR models, the customized BERT model BETO was used, which was a model pre-trained on an unlabelled Spanish dataset. Parameters and hyperparameters of this model were

fine-tuned separately for the tasks, i.e. binary classification and regression. The models were trained through three unique scenarios, where the data fed to the models were different for each scenario. The metrics that were used to evaluate the performance of the models in [21] on each task were different. The quality and performance of the binary classification model were evaluated using AUROC, Weighted Accuracy, Weighted Recall and Weighted F1 Score; whereas those of the LR model were evaluated using the metrics MSE, Mean Absolute Error (MAE), and Coefficient of Determination (R2). For the binary classification task, although the model that was trained on scenario 2 ended up having the highest performance score, it was discarded since there was a bias within the database, resulting in the model that was trained on the dataset of scenario 3 to be declared the best performing one. Most of the errors that were made by this model were due to the preprocessing pipeline, which ended up removing some meaningful information from the dataset and hence necessitated the need for a baseline for the quality of messages so that the model could identify faulty messages successfully. Similarly, for the LR task, the model trained on scenario 2 was discarded, regardless of it achieving first place in terms of performance, due to the dataset having a bias. Thus, the model trained on scenario 3 was deemed to be the top performer. Additionally, the dataset here lacked useful information too, therefore, it also required a threshold to be fixed for the quality of messages.

Recently, in the paper [20], the authors (Guerra et al.) aimed to prove that classical machine learning methods could perform better than transformers in simple NLP-related tasks. They (Guerra et al.) tried doing this by experimenting with the MentalRiskES task of the IberLEF (Iberian Languages Evaluation Forum) shared task. The two tasks that were focused on were: firstly, identifying eating disorders (binary classification) and the probability of anorexia or bulimia (simple regression); secondly, detecting depression and estimating its probability (simple regression), and classifying it into a fixed set of classes (multi-output regression). The tasks were implemented on the dataset given by the MentalRiskES from the IberLEF 2023 challenge, which consisted of 50 rounds of Telegram messages for the first task and 100 rounds of Telegram messages for the second. After splitting the dataset into train and test data, the authors (Guerra et al.) implemented preprocessing methods, which included using TF-IDF to extract features, and it was done with bigrams to conserve the context. For the binary classification task, the machine learning models used were SVC, MNB, KNN, and Extreme Gradient Boosting (XGB), with all having the pre-defined parameters of the scikit-learn library except MNB, where a slight change was made. For regression tasks, Stochastic Gradient Descent Regressor (SGDRegressor), Ridge Regressor, Linear Regressor, Gradient Boost Regressor Trees (GBRT), and Random Forest Regressor were used. The model with the best performance for each task was selected and was then compared with the baseline transformer models of the MentalRiskES tasks - RoBERTa-Base, RoBERTa-Large and DeBERTa. In this study [20], the evaluation perspectives considered for the classification tasks were absolute classification and early detection effectiveness, for which Macro F1, and ERDE, in this case ERDE30, were used respectively. For the regression tasks, Root Mean Square Error (RMSE) and Precision@k (P@k), in this case P@30, were used for the evaluation perspectives of error basis and ranking basis respectively. Overfitting problems were the primary sources of errors, where

the models were unable to comprehend the context of the messages. The results obtained after solving the tasks indicated that classical machine learning models could perform better than the deep learning architecture of transformers in the absolute classification and early detection tasks, provided no problem occurred during the training phase. However, in simple regression and more complex tasks like multi-output regression, transformers surpassed the traditional machine learning models in terms of performance metrics since they were pre-trained on a huge corpus, which proved to be a great advantage in such tasks. Nevertheless, if the machine learning models are trained on a larger dataset, they are expected to perform at the same level as the transformers, or maybe even outperform them.

Chapter 3

Dataset

3.1 Analysis of Dataset

We collected our dataset from CLEF eRisk 2024 Task2: Early Detection of Signs of Anorexia, which was a continuance of eRisk 2018's T2 and eRisk 2019's T1 tasks [24], and we used the train and test data from 2018's T2 task. The dataset consisted of writings, more precisely, it was a collection of posts or comments from social media users where the users were divided into two categories: anorexic or non-anorexic (the control group). As the task was for early detection, the writings were sequenced in chronological order for each user. The dataset had 10 chunks for both positive (anorexia) and negative (control), where each chunk consisted of the writings of a set of users, and the chunks were sequenced in chronological order, for example, chunk 1 may have consisted of users' writings from a specific period (e.g., 2001-2002), while chunk 2 contained the subsequent writings (e.g., 2003-2004), and so on. The positive and negative data of the train set consisted of different sets

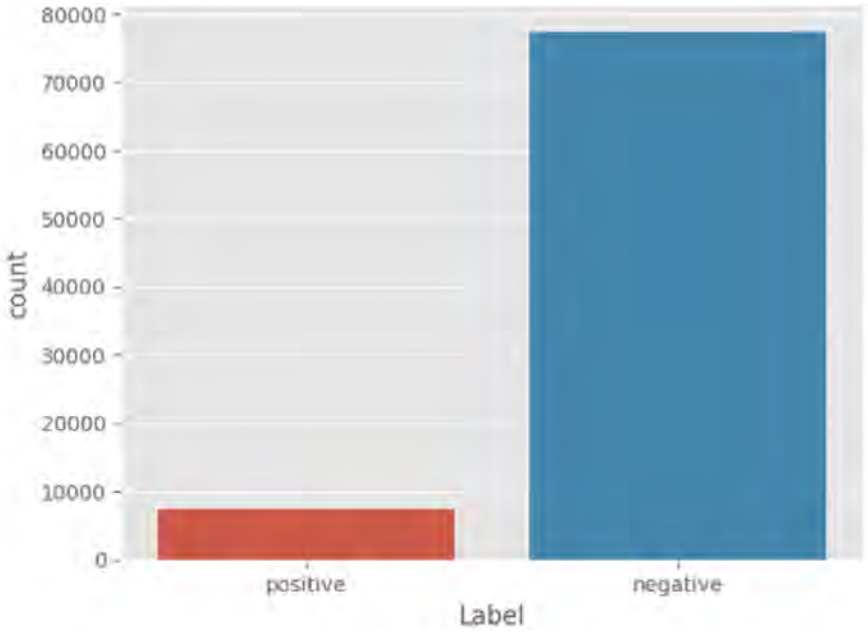

Figure 3.1: Row Distribution

of users (subjects), to be more precise, 20 subjects and 132 subjects respectively,

while the test data had an overall 320 subjects. The positive dataset of the training data consisted of 7,452 rows and 5 columns, and the negative consisted of 77,389 rows and 5 columns. The row distribution is shown in Figure 3.1. After merging the positive and negative data of the training dataset into a single dataframe, the dataset accumulated to 84,841 rows and 5 columns. The distribution of the dataset is further clarified in Figure 3.2.

	eRisk 2018 - Anorexia Task			
	Anorexia	*Control*	*Anorexia*	*Control*
Num. subjects	20	132	41	279
Num. submissions (posts & comments)	7,452	77,514	17,422	151,364
Avg num. of submissions per subject	372.6	587.2	424.9	542.5
Avg num. of days from first to last submission	803.3	641.5	798.9	670.6
Avg num. words per submission	41.2	20.9	35.7	20.9

Figure 3.2: CLEF eRisk 2018 Anorexia Dataset Statistics

	ID	TITLE	DATE	INFO	TEXT
0	subject758	NaN	2013-10-09 02:26:19	reddit post	I googled her and ha, that's not me..but inte...
1	NaN	NaN	2013-10-08 15:08:59	reddit post	in terms of book genres- fiction, surrealism/...
2	NaN	NaN	2013-10-08 14:45:07	reddit post	Breaking Bad, I suppose. I read a lot of Mura...
3	NaN	I'm a 19 (F) and I guess this is my story. De...	2013-10-08 14:07:15	reddit post	I'm a 19 year old girl. \n\n I'm mentally ...
4	NaN	NaN	2013-10-07 12:17:02	reddit post	two experiences. \n\nMy grandma and auntie ha...
5	NaN	NaN	2013-10-06 06:30:16	reddit post	my sister and I walked in on our grandpas car...
6	NaN	NaN	2013-10-02 13:42:52	reddit post	crying.
7	NaN	NaN	2013-10-01 07:14:50	reddit post	wtFFuuHHuhaskjaf. ???
8	NaN	NaN	2013-10-01 05:47:06	reddit post	Ridiculously photogenic male bungee jumps in ...
9	NaN	NaN	2013-09-30 13:37:35	reddit post	"I bet she fucks like a fairy on acid."

Figure 3.3: First 10 Rows of Combined Dataframe After Changing The Column Names

We divided the training dataset into two parts: training and validation, with the train set having 15 positive anorexia users and 106 negative anorexia users, and the validation set having 5 and 26 anorexia-positive and anorexia-negative subjects respectively.

The columns of both the train and test dataset consisted of the subject ID (ID), the title of the post/comment (WRITING___TITLE), the time the writing was posted in the format YYYY-MM-DD hh:mm:ss (WRITING___DATE), further information about the post (WRITING___INFO), and the text writing itself (WRITING___TEXT). We further divided the WRITING___TEXT column into categories

based on the number of words per post. Comparing the positive and negative datasets, it can be seen that the majority of the posts were short in both datasets, with around 3,750 short texts in the positive data and 57,500 in the negative. While the number of very large posts was insignificant in the negative dataset, in the positive dataset the frequencies of medium-sized posts and very large-sized posts were almost identical. The distribution displayed in Figure 3.4 primarily illustrates the dominant usage of concise posts in both datasets.

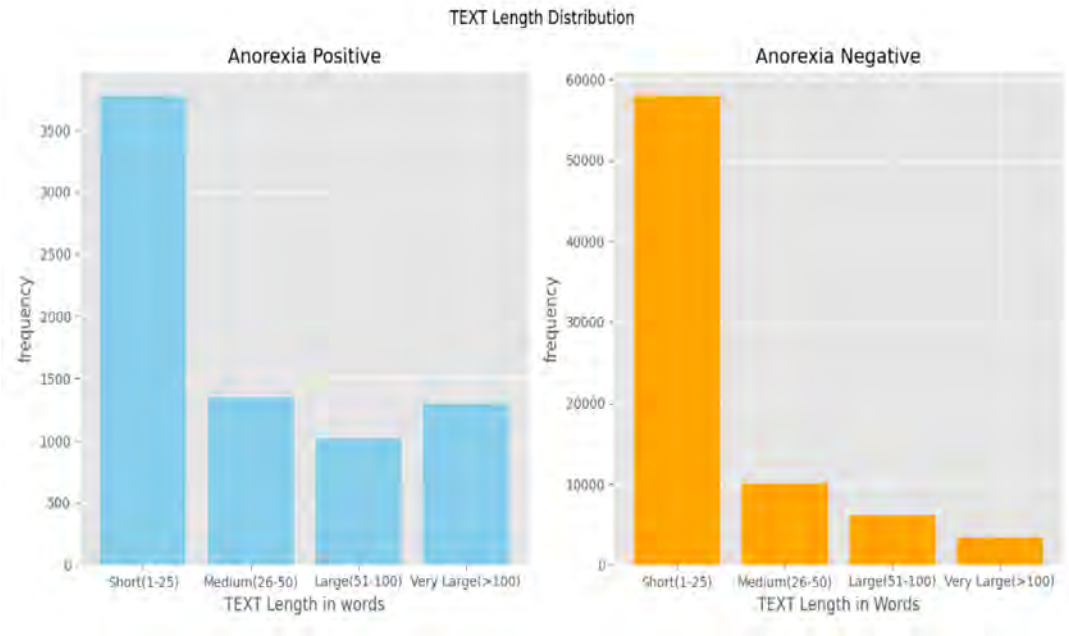

Figure 3.4: TEXT Column's Row Length Distribution

Both the positive and negative datasets had many URLs in the posts. Apart from the URLs, the most common word in both the datasets was **like**, while words like **want**, **discussion**, and **eating** were among the top ten common words of the anorexia-positive dataset, having similar frequencies of approximately 1600, 1225, 1200, and 1200 respectively.

3.2 Data Preprocessing

The dataset that we collected from CLEF eRisk 2024 Task2 was a collection of files that were all in XML form. Due to this format being outdated and less used in recent times, we decided to convert all of the XML files into the more popularly used CSV format. For this, we used the online "XML to CSV Converter" [29] tool. Next, we separated the CSV files subject-wise, e.g. a folder named subject38 contained all 10 chunks of an anorexia-positive subject - subject38's CSV files, and we did this for all the positive and negative subjects. We then merged all 10 chunks of an individual subject into a single CSV file and proceeded to do this with all the subjects. Finally, we combined all the positive and negative subjects' CSV files so that ultimately we were left with just two CSV files: positive (merged files of all the anorexia-positive

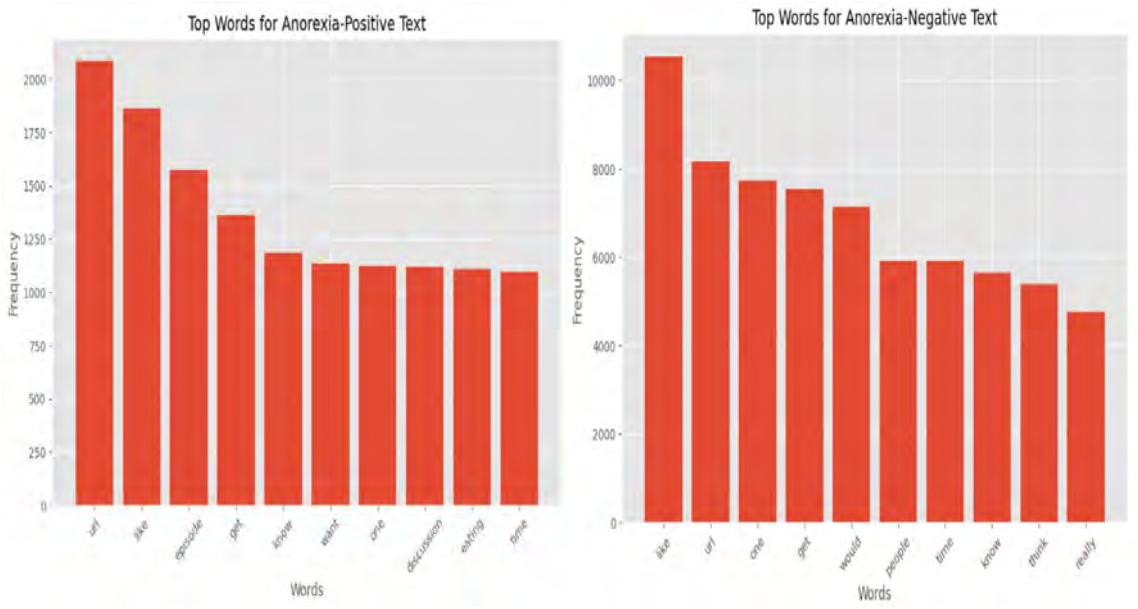

Figure 3.5: Top 10 Words In Anorexia-Positive and Anorexia-Negative Datasets

subjects) and negative (merged files of all the anorexia-negative subjects). We used merge-csv.com [27], which is an online tool for merging CSV files.

After attaining the combined positive and negative datasets, we merged these two into a single dataframe. There were many empty rows in the WRITING___TEXT column and to reduce these, we added the row values (if any) in the WRITING___TITLE column to the corresponding WRITING___TEXT row values. Then, we removed the remaining few null values to eliminate empty rows in the WRITING___TEXT column. Next, we applied a series of preprocessing methods on the posts in the WRITING___TEXT column:
First, we replaced all the URLs and HTML tags with the placeholder <url> and <html> respectively, since these were in abundance in our dataset but they normally do not contribute to NLP contexts. We also removed punctuations, usernames, newline characters, and other garbage characters for the same reason. Emojis play an important role in describing moods and due to their high usage in day-to-day texts, they are very significant in adding context to NLP-related data. Hence, to conserve the sentimental meaning and expressions from emojis in our dataset, we addressed their presence by converting them to their corresponding semantic textual representation using Python's emoji module, e.g. 😢 is replaced with :cry:. Additionally, we expanded contractions like **lol**, **rofl**, **gn**, and so on, to their actual forms - **laugh out loud**, **rolling on the floor laughing**, and **good night** respectively. For this, we created a dictionary of around 100 contractions and their expanded forms, which we thought were popularly used in social media texts, This step was important because expanding these contractions can provide more semantic clarity and context, which may in turn be crucial for accurately identifying and understanding the subtleties of language that might indicate anorexia. After that, we separated the posts of all the users into sub-dataframes and kept these in a dictionary, where the key was the dataframe containing all the posts of a user, and the value was the corresponding

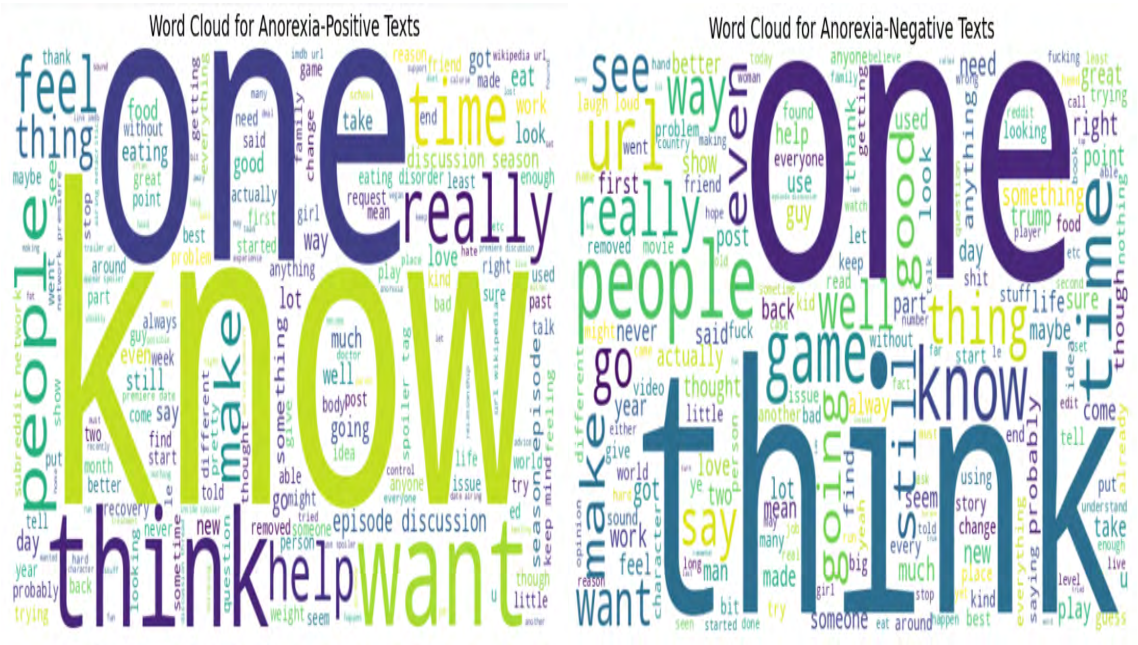

Figure 3.6: Word Clouds for Anorexia-Positive and Anorexia-Negative Datasets

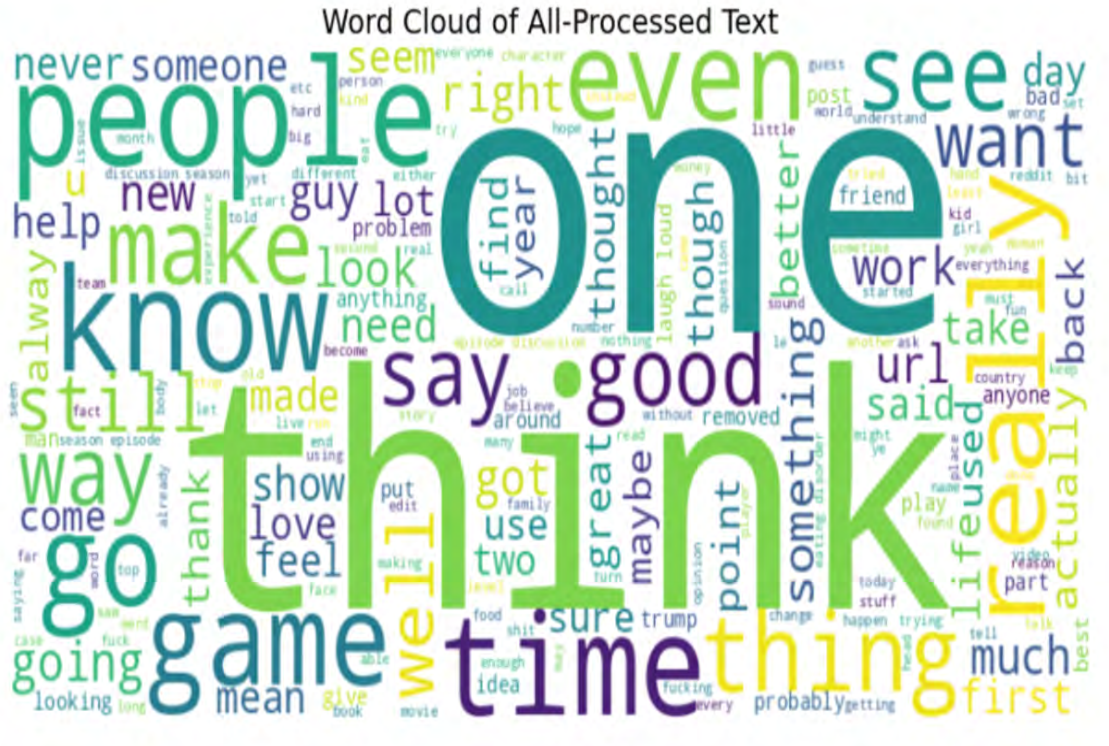

Figure 3.7: Word Cloud For Combined Dataset After Preprocessing

label of the user, i.e. whether the user was anorexia-positive or anorexia-negative. Furthermore, we split the data into two parts randomly using Python's random module, 80% of the users were allotted to the training set, and the remaining 20% were allocated to the validation set. Due to the significant imbalance in positive and negative anorexia users in our dataset, we ensured that the validation set consisted

of 20% users of each class from the original list of users. This specific splitting of the data ensured that there was enough data for training purposes, and also that a reasonable portion of the data was left to be used for validation processes. Finally, during both training and testing, we used BERT to generate sentence embeddings of all the posts of each user, before the entire set of embeddings per user was fed to our models.

Chapter 4

Model

4.1 Long Short-Term Memory (LSTM)

An LSTM (Figure 4.1) is a category of RNN, however, it was built in such a way that it could overcome the shortcomings of classical RNNs in recording long-term dependencies in sequential data. It can do so due to the presence of memory cells and gates that enable specific information to be conserved and forgotten over time. Delving deeper into an LSTM's structure, information can be retained for long durations since there is an internal memory state, and this is what allows it to capture long-term dependencies.

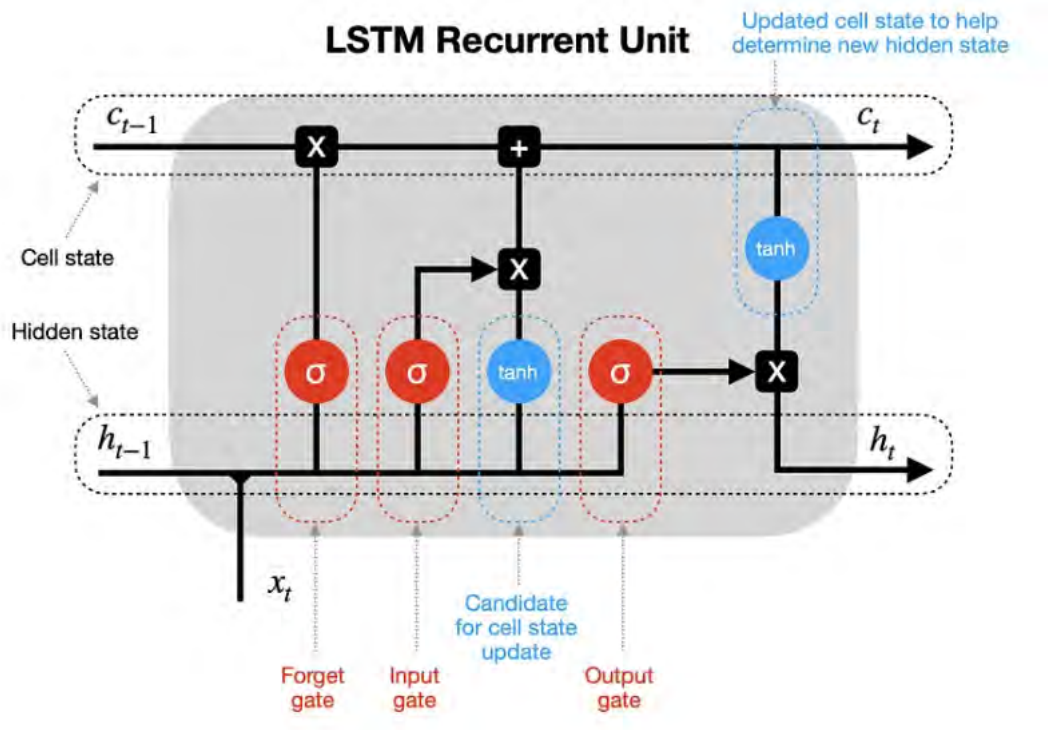

Figure 4.1: A single LSTM Recurrent Unit [23]

4.1.1 Bi-Directional Long Short-Term Memory (Bi-LSTM)

Another class of RNNs is Bi-LSTM (Figure 4.2) which possesses the ability to not only process sequential data in the forward direction but also the backward direction. The Bi-LSTM model is more powerful than its simpler version, LSTM, in the sense that it incorporates the features of the LSTM with bidirectional processing, allowing the context of both the past (i.e. previous time steps) and future (i.e. following time steps) of an input sequence to be captured. A Bi-LSTM model has two layers of this LSTM structure, one processes the input sequence in the forward direction while the other does so in the opposite direction, and the memory cells and hidden states are maintained by each layer. The processing of the input sequence is further described below:

1. From the first time step to the last, the forward LSTM layer receives the input sequence during the forward pass, where at each time step it evaluates its hidden state and modifies its memory cell based on the previous memory cell and hidden states and the current input [18].

2. The backward pass occurs at the same time as the forward pass and the process is identical to the forward pass except for one part, the input sequence is fed into the backward LSTM layer in reverse order, and from the last time step to the first [18].

3. Once these are finished, the hidden states from both the LSTM layers are merged at each time step and this integration can be a simple concatenation operation of the hidden states or some other transformation [18].

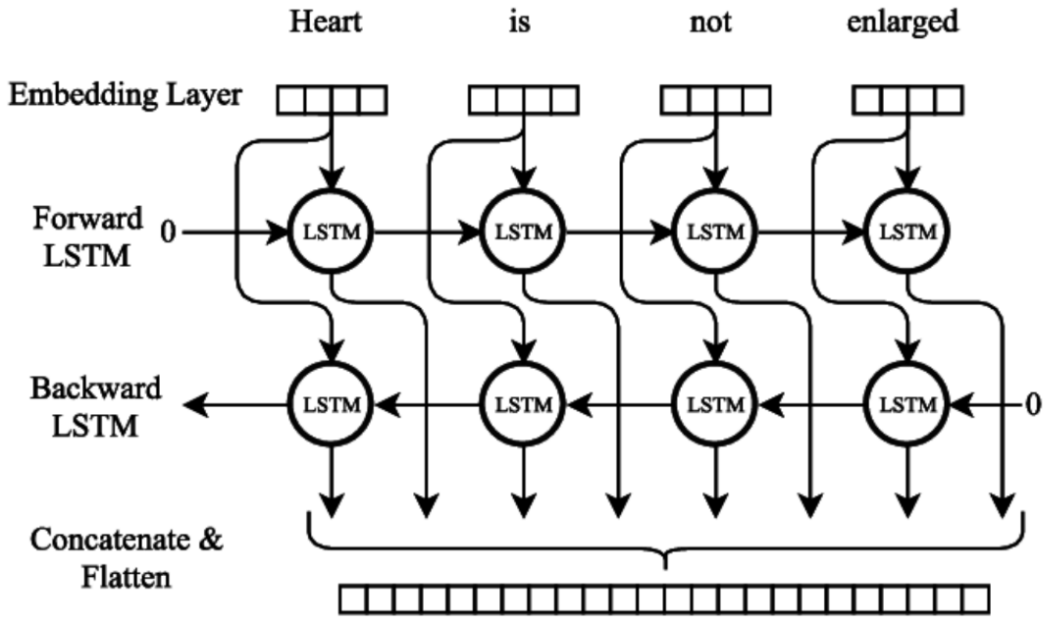

Figure 4.2: Bi-LSTM Architecture [1]

By leveraging information from both directions, Bi-LSTMs generally offer better performance and enhanced accuracy compared to unidirectional LSTMs in various applications. They are hence more versatile and are suitable for a wide range of

tasks, including POS tagging, named entity recognition (NER), machine translation, text classification, and more. To summarize, the bidirectional nature of Bi-LSTMs makes them powerful tools for understanding and grasping full context in sequential data, leading to improved model performance and accuracy. Considering all these advantages of Bi-LSTM, we decided to implement text classification on our dataset using the aforementioned model.

4.2 Gated Recurrent Unit (GRU)

GRU is another RNN variant, closely related to the LSTM network. Both are designed to work with sequential data, enabling the network to selectively retain or discard information over time. However, the GRU is more streamlined in design, with fewer parameters, making it quicker to train and less computationally demanding compared to LSTM. The major difference between GRU and LSTM is how they handle memory. In LSTM, the memory cell state is maintained separately from the hidden state, and its updates are controlled by three gates: the input gate, the output gate, and the forget gate. In contrast, GRU replaces the memory cell state with a "candidate activation vector" and simplifies the process by using only two gates: the reset gate and the update gate. The reset gate decides how much of the past hidden state to erase, whereas how much of the candidate activation vector is integrated into the new hidden state is determined by the update gate[19]. Figure 4.3 shows a single GRU unit.

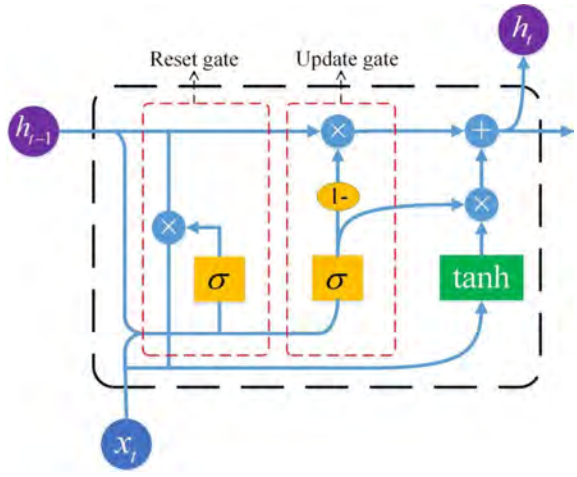

Figure 4.3: A Single GRU Unit [11]

4.2.1 Bi-Directional Gated Recurrent Unit (BiGRU)

A BiGRU processes input sequences in two directions: forward (from start to finish) and backward (from end to start). Unlike a standard GRU, which only looks at the past context of the sequence, a BiGRU captures both the past and the future information[22]. This is achieved by maintaining two separate hidden states, one for each direction, and combining the hidden states from both before making a final

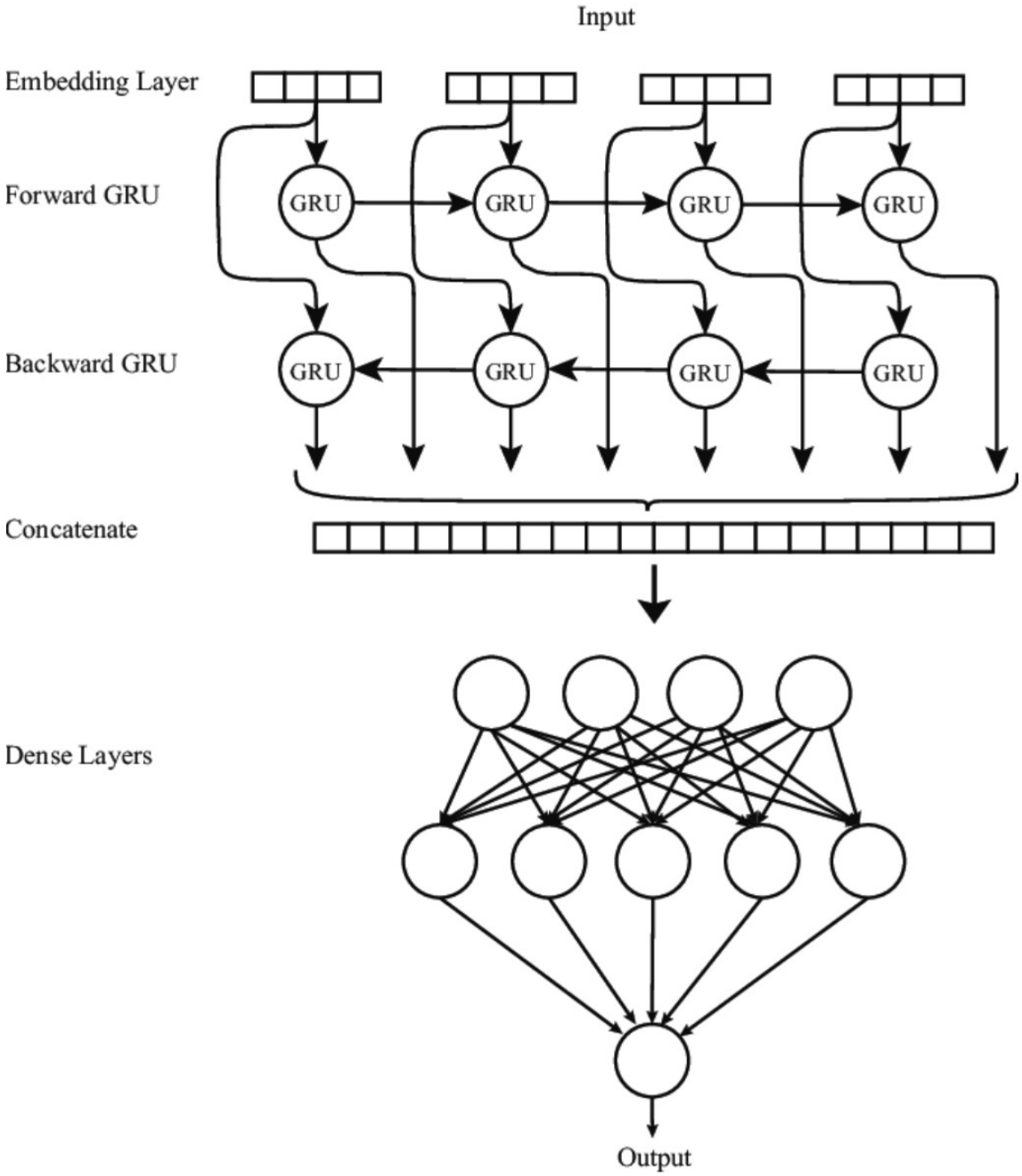

Figure 4.4: BiGRU Architecture [12]

prediction. By utilizing this bidirectional approach, BiGRUs are able to learn patterns in the data that may not be apparent when only past information is available. The model can more effectively understand the relationships between elements at different points in the sequence. This method is particularly useful in tasks where the entire context of a sequence is important, such as in natural language processing or time series analysis.

4.3 BERT

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a model in the field of Machine Learning designed for NLP. Created by researchers at Google AI Language in 2018, it functions as a versatile tool that addresses over 11 common language tasks, including sentiment analysis and named entity recognition. Historically, enabling computers to truly "understand" language has been a challenge. While machines can collect, store, and interpret text inputs, they often struggle with the nuances of language context. This is where NLP comes in—an area of artificial intelligence dedicated to enabling computers to read, analyze, interpret, and derive meaning from both text and spoken words. NLP merges elements of linguistics, statistics, and Machine Learning to help machines better grasp human language. In the past, each specific NLP task was tackled with dedicated models. However, BERT has changed that paradigm! By effectively addressing over 11 prevalent NLP tasks, BERT has emerged as a leading model in the field, outperforming many of its predecessors and establishing itself as a multifaceted tool for NLP applications. BERT's success can be attributed to its training on a vast dataset containing 3.3 billion words. It was primarily trained using Wikipedia, which accounts for approximately 2.5 billion words, and Google's BooksCorpus, contributing around 800 million words. These expansive datasets have enriched BERT's understanding of both the English language and the world at large! Training a model of this scale requires significant time investment. BERT's development was facilitated by the innovative Transformer architecture and accelerated through the use of TPUs (Tensor Processing Units), which are specialized circuits developed by Google for large-scale ML tasks. Remarkably, BERT was trained using 64 TPUs over a span of just four days.

4.3.1 BERT's Mechanism

The key to BERT's effectiveness lies in its use of the Transformer architecture[3], which allows it to analyze the context of words bidirectionally. Earlier models, like LSTMs, typically processed text either from left to right or right to left, meaning they could only focus on either the preceding or following words. In contrast, BERT reads text in both directions at once, enabling it to fully grasp the meaning of each word by looking at both the words before and after it. BERT's architecture involves two main steps: pre-training and fine-tuning.

1. **Pre-training**: During this phase, BERT was trained on large datasets like Wikipedia and BooksCorpus. Two major tasks guided this training:
    - **Masked Language Modeling (MLM)**: In this task, some words in a sentence were masked, and BERT had to predict the missing words based on the surrounding context. This helped the model learn better word relationships than traditional next-word prediction methods.
    - **Next Sentence Prediction (NSP)**: This task trained BERT to understand the relationship between sentences. The model was given two sentences and it had to predict whether the second sentence logically fol-

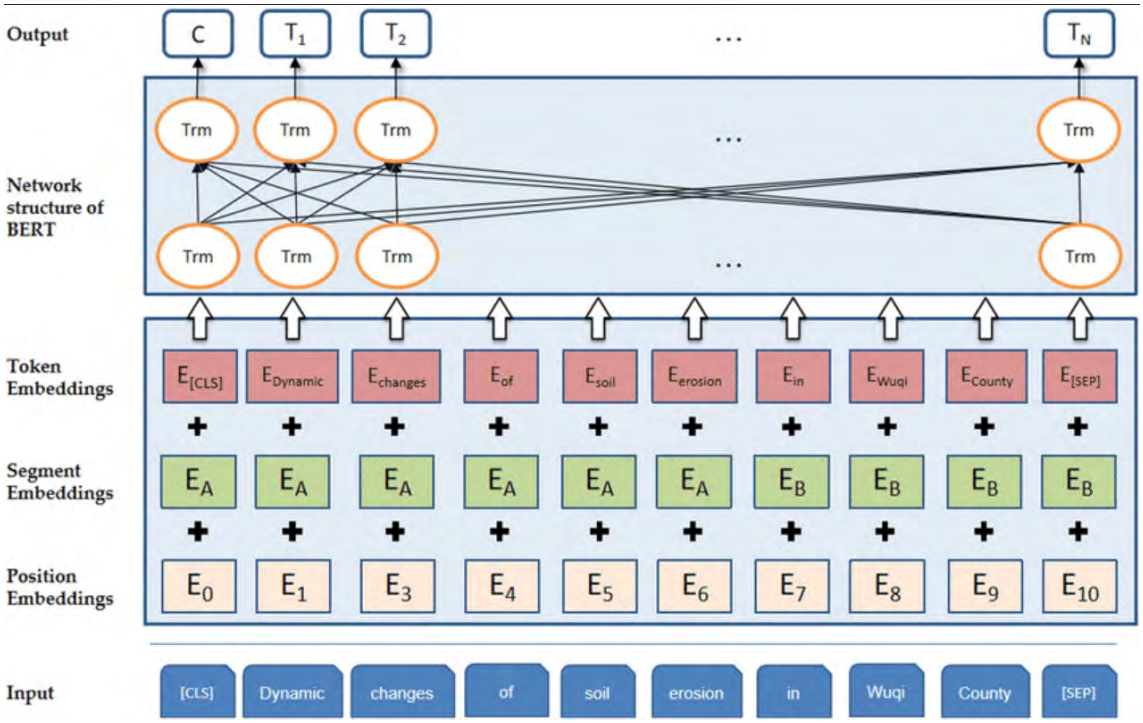

Figure 4.5: How BERT Works [16]

lowed the first. This helps in tasks that require understanding of multiple sentences, such as question-answering and entailment tasks.

2. **Fine-tuning**: After pre-training, BERT is adapted to specific NLP tasks by fine-tuning it on smaller datasets. Fine-tuning requires minimal adjustments because BERT already has a deep understanding of language patterns, thanks to the pre-training phase. This allows BERT to be applied effectively to a wide range of tasks, such as text classification, question-answering, and more.

4.3.2 BERT Encoder: Input Dimensions and Processing

The BERT model specifically utilizes the transformer model's encoder portion. It is designed to generate rich, contextualized representations of text sequences. Below is a breakdown of its inner workings and input dimensions.

Input Representation

BERT processes a sequence of tokens, which may represent sentences or sentence pairs (as in tasks like question answering).
Before feeding data into the BERT model, the input text undergoes tokenization using the BERT tokenizer. This tokenizer divides the text into subword tokens and converts them into numerical IDs that correspond to entries in BERT's pre-trained vocabulary.
For example, the sentence `"I love judo"` might be tokenized as follows:

$$['[CLS]', 'I', 'love', 'ju', '##do', '[SEP]']$$

Here, `'[CLS]'` is a special token indicating the start of the sequence, and `'[SEP]'` is a separator used to distinguish between different sentences in tasks that involve

pairs of sentences. Subword tokens (like `'ju'` and `'##do'`) are a feature of the BERT tokenizer.

BERT Input Components

The inputs fed into BERT consist of the following:

- **Input IDs**: These are the tokenized IDs from the BERT vocabulary.
- **Attention Mask**: A binary mask that indicates which tokens in the sequence are actual words (1) and which are padding tokens (0).
- **Token Type IDs**: Used when processing sentence pairs to differentiate between the two sentences. For single sentence tasks, this is usually set to all 0s.

Embedding Layers

BERT converts the tokenized input into embeddings through the following three components:

- **Token Embeddings**: Each word token is mapped to a fixed-size vector.
- **Position Embeddings**: Since the transformer architecture lacks inherent sequential information, position embeddings are added to indicate the position of tokens within the sequence.
- **Segment Embeddings**: These help differentiate between two different sentences in sentence-pair tasks.

The resulting embeddings for each token form a tensor of shape `(batch_size, sequence_length, embedding_size)`. For BERT-base, the `embedding_size` is 768, while for BERT-large, it is 1024.

BERT Encoder Architecture

The BERT encoder, as illustrated in Figure 4.6, comprises a stack of transformer encoder layers, each applying self-attention followed by feed-forward neural networks.

- **Self-Attention Mechanism**: This allows each token in the input to attend to other tokens in the sequence, considering contextual information from both directions (i.e., left and right of the current token).
- **Multi-Head Attention**: BERT uses multiple attention heads, enabling it to focus on different parts of the sequence simultaneously.
- **Feed-Forward Networks**: Following the attention mechanism, fully connected layers (feed-forward networks) further process the information.

The output from each transformer layer maintains the same shape as the input, i.e., `(batch_size, sequence_length, embedding_size)`.

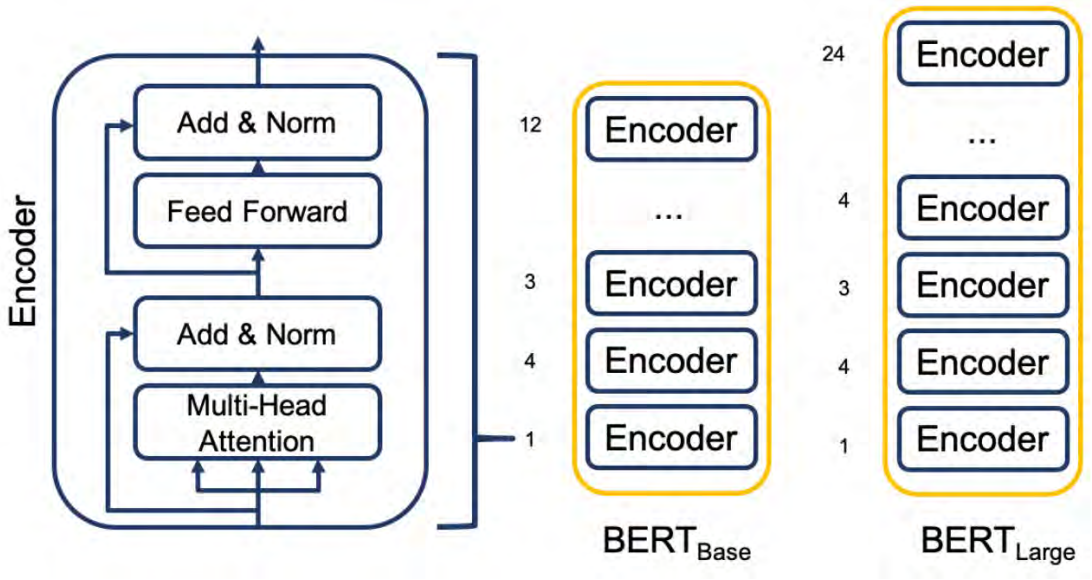

Figure 4.6: BERT Base and BERT Large Encoder Architecture [25]

Output from BERT Encoder

The final output from the BERT encoder is a sequence of embeddings, one for each token in the input, incorporating context from both directions. The output tensor has the shape:

```
(batch_size, sequence_length, hidden_size)
```

For BERT-base, the `hidden_size` is 768, while for BERT-large, it is 1024.
For tasks involving classification (e.g., sentiment analysis), the output corresponding to the special `'[CLS]'` token (the first token in the sequence) is often extracted. This yields a tensor of shape `(batch_size, hidden_size)`, where the hidden size is typically 768 or 1024 depending on the BERT variant.
For token-level tasks (e.g., named entity recognition), the full sequence of token embeddings are utilized, giving an output of shape `(batch_size, sequence_length, hidden_size)`.

BERT's ability to consider both the preceding and following context in a sentence makes it especially good at understanding the true meaning of words that might have multiple meanings. For example, in the sentence "The pitcher threw a strike," BERT can correctly identify whether "pitcher" refers to a person throwing a baseball or a container used to pour liquids, depending on the surrounding words.

Additionally, BERT can handle inputs of varying lengths and complexities, which makes it highly flexible for different NLP tasks. Its combination of bidirectional context and Transformer-based architecture allows it to capture subtle nuances in text, making it one of the most powerful tools for language understanding today. This unique way of processing context and relationships between sentences is what sets BERT apart from earlier models and continues to make it a significant advancement in the world of NLP.

After manually tuning our model's hyperparameters to find the best results, we

ended up with 128 hidden dimensions, a learning rate of 0.00005, and a dropout probability of 0.5. For our optimizer, we chose Adam, since it is widely regarded as the best for NLP related tasks. Additionally, we used weight decay of 1e-5 for our models and trained each model for 30 epochs, with early stopping to avoid overfitting. For the LSTM and GRU models we set the patience to 3, whereas for BiLSTM and BiGRU, we fine-tuned it to 2.

Chapter 5

Experiment and Result Analysis

In this study, we aimed to evaluate the performances of LSTM, BiLSTM, GRU and BiGRU as binary classification models for the early detection of anorexia on our dataset, which is mainly composed of textual data of social media users whose posts may or may not show signs of anorexia. To train and evaluate our models, the first thing we did was divide the data into two parts: 80% for training and 20% for validation. For BERT's bidirectional context capturing ability from a text sequence, for each subject we used it to produce contextualized post embeddings for all their posts and then fed these post embeddings together along with the corresponding label to our model during the training phase. We used the pre-trained "bert-base-uncased" model and its correspondent tokenizer. The process is visualized in Figure 5.1.

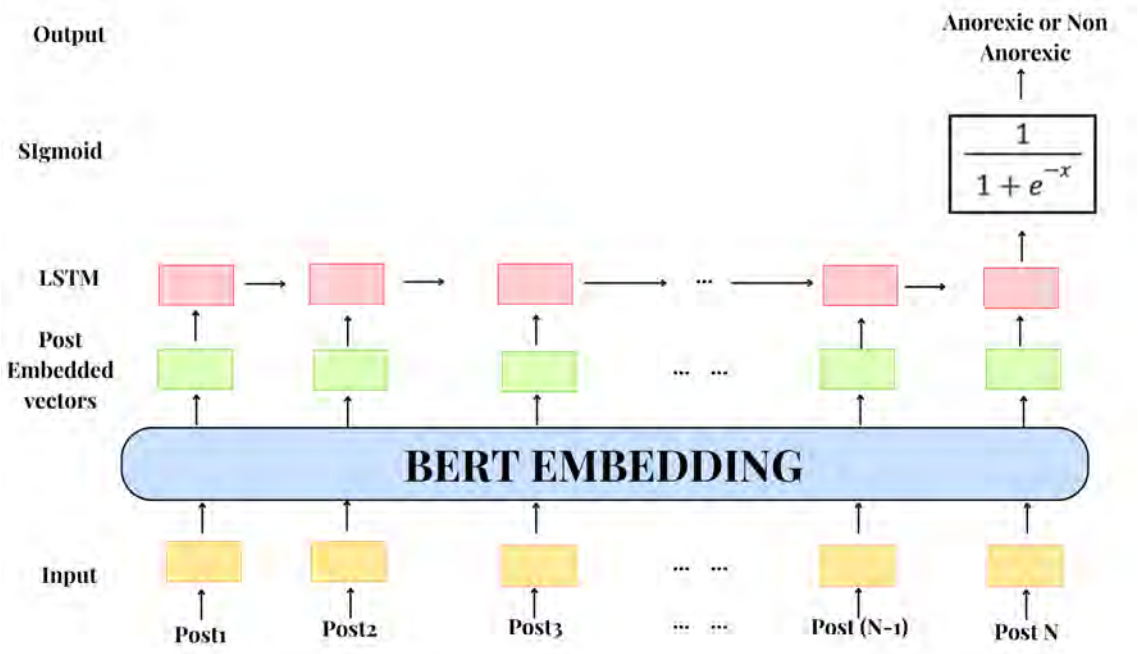

Figure 5.1: Using BERT To Generate Contextualized Post Embeddings and Feed It To LSTM

We set the tokenizer sequence length to a maximum length of 512, truncation and

return_attention_mask were set to True, and we also set the padding equal to the maximum sequence length. We sent the posts of a subject to the tokenizer in batches of 16 for better efficiency, and the tokenizer returned the post embeddings as Pytorch tensors. These embeddings were appended in a Python list and were eventually stacked together using Python's NumPy module.

As mentioned before, our training dataset was greatly imbalanced, and hence, to overcome this obstacle we used the weighted binary cross-entropy loss as our cost function. The anorexia-positive subjects were given a weight equal to the ratio of anorexia-negative subjects to the total number of subjects, and the anorexia-negative subjects were given a weight equal to the ratio of anorexia-positive subjects to the total number of subjects. This ensured that the anorexia-positive subjects, which belonged to the minority class, received a greater weight compared to the anorexia negative subjects. By assigning a higher weight to the minority class (positives), the model was essentially told that misclassifying a positive sample was more costly than mislabelling a negative one. This encouraged the model to focus more on correctly classifying the positive samples. This resulted in a great boost and faster convergence in the training performance of all the models.

In addition to this, we also shuffled the training data subjects at the beginning of each epoch, and ensured that not more than two anorexia-positive subjects could be fed to the model in succession. This method helped in bringing randomness to the order of the training data, which was originally sorted with the positive labels coming first followed by the negative ones, and it was necessary to undergo this process since data in the real world cannot be found in a sorted arrangement. Moreover, having this random arrangement helped improve the performance and mitigate model overfitting as well.

Coming to the testing and early detection part, we executed the same method of generating post embeddings as we did during the training phase, however, instead of producing embeddings for all the posts of a user at once, we both created the embeddings using BERT and then fed these to the model chunk by chunk, i.e in the original form of the testing data instead of concatenating all the chunks into a single file. If a chunk was not classified as anorexic, we merged it with the next chunk and fed the total post embeddings to the model and this was continued till the model being used detected anorexia, or all the chunks were used up. Diving a little deeper into our classification and early detection approach, we implemented a two-step system. We checked at first whether or not the probability score of the user being anorexic up to all the chunks being fed was greater than 70%. If true, we considered the user to have a good likelihood of having anorexia and concluded them to be anorexic, and this was our threshold for early detection. If the probability score was less than the threshold of 70%, we performed an additional check, inspired by the approach used in the paper [4], we built a detection threshold mechanism. When the prediction value till the number of chunks fed to the models came within the range of 50%-70%(exclusive), a flag was raised and a secondary evaluation phase was entered where the model set a threshold of three chunks, i.e. if the prediction score came in the given range again within the inclusion of the next three chunks, the user was categorized as having anorexia, else the flag was set down and the

testing period of the user continued. This is better visualized and is clarified in Figure 5.2. We regarded the extra evaluation measure vital for reliable classification as a user who may have shown signs of anorexia to some degree in the earlier chunks of their total data may have eventually turned out to be non-anorexic. Finally, if no decision could be made until the last chunk, the last probability value was used for the ultimate outcome, where a score higher than or equivalent to 50% was taken as anorexic, and vice versa.

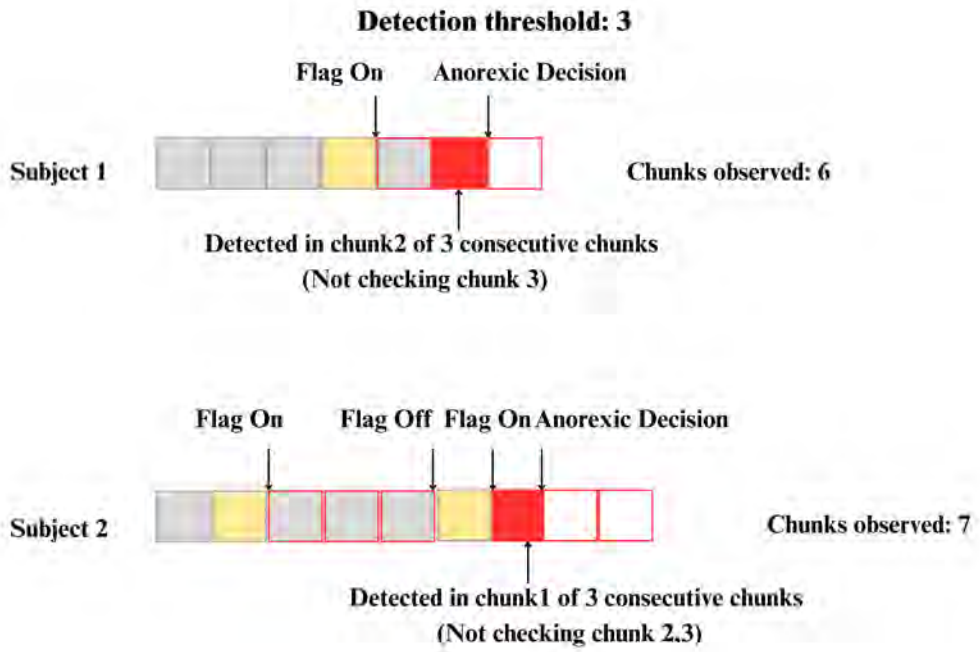

Figure 5.2: Detection Threshold Mechanism

5.1 LSTM Training Evaluation

In the training part from Figure 5.3, we can observe that both training and validation loss was decreasing at first from training loss at 15.03% and validation loss at 17.3% but after a few epochs the validation loss curve started to fluctuate which resulted in the early stopping at 14 epochs with the training loss of 4.37% and the validation loss of 13.9% approximately.

5.2 BiLSTM Training Evaluation

During training, Bi-LSTM performed well on the training data but had some difficulty predicting the validation data, as can be understood from Figure 5.4. The training loss decreased from around 15% to a little over 6%, while the validation loss dropped to a minimum of approximately 6.5%. Although the general trend of

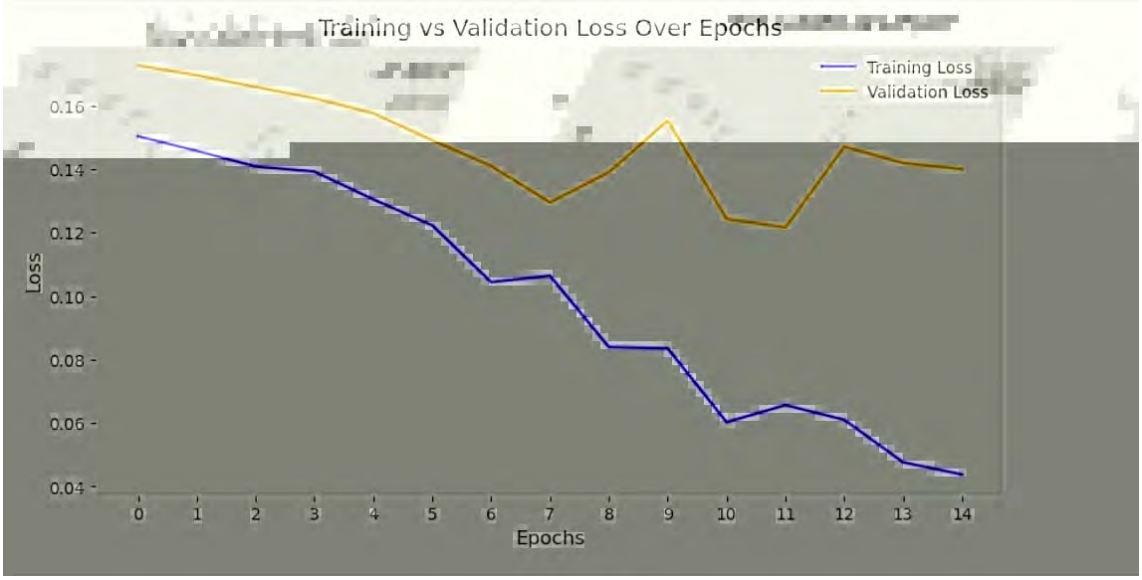

Figure 5.3: Training and Validation Loss of LSTM Over Epochs

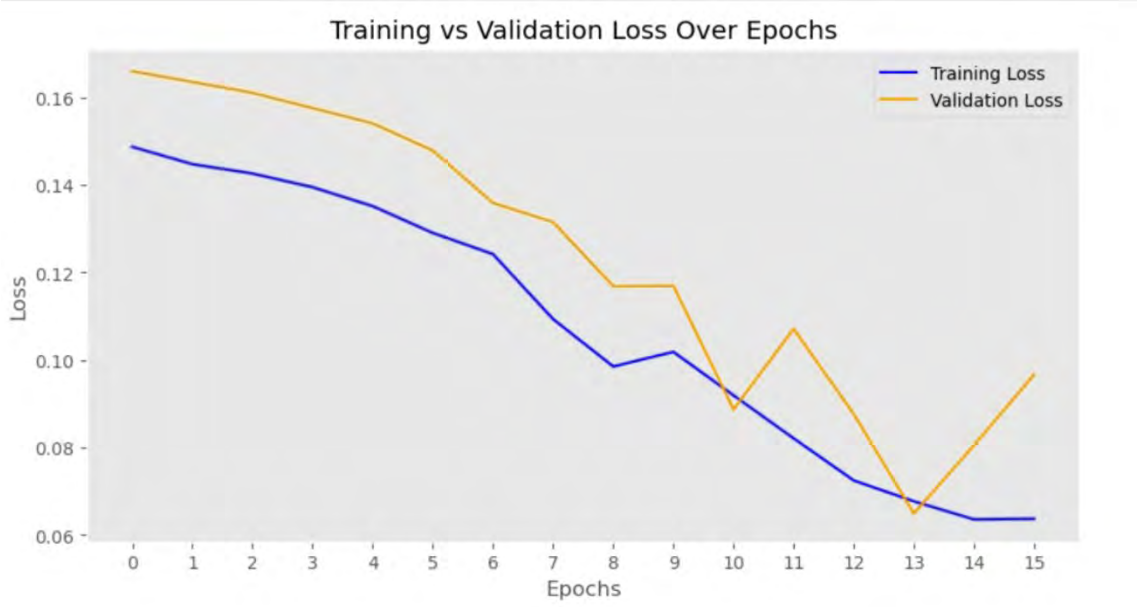

Figure 5.4: Training and Validation Loss of BiLstm Over Epochs

both the losses seem to be decreasing over the epochs, the validation loss was a bit higher than the training loss, and it also started greatly fluctuating from the 10th epoch, which resulted in the early stopping of the model. These indicate that the model was able to capture underlying complex patterns from the data, but it also fit on the datapoints too well, which ended up in the model being overfitted and early stopping being executed.

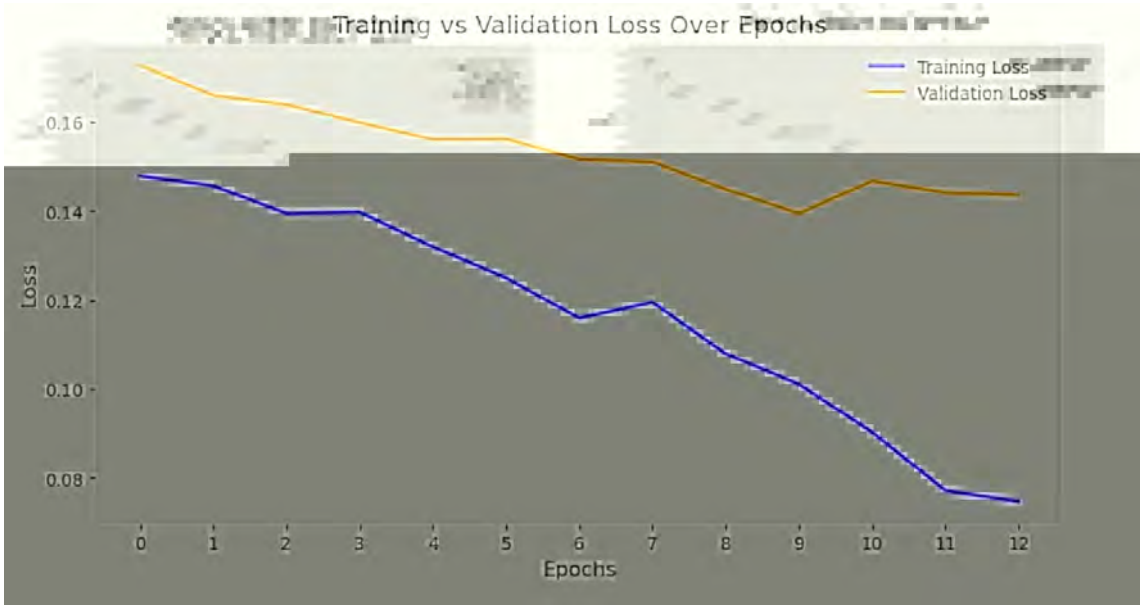

Figure 5.5: Training and Validation Loss of GRU Over Epochs

5.3 GRU Training Evaluation

Initially, both training and validation losses were at 14.35% and 16.78% and the losses were decreasing smoothly as we can observe in Figure 5.5 . However, after the 9th epoch, the validation loss curve began to change rapidly depicting an increasing trend. This increasing trend implies that the model has already succeeded in understanding the underlying patterns in the data, and now it may start to overfit. Ultimately, resulting in an early stop at 12 epochs with a training loss of 6.5% and a validation loss of 14.7%.

5.4 BiGRU Training Evaluation

BiGRU performed well on training data but struggled to predict validation data, as seen in Figure 5.6. The training loss fell from 16% to 4% approximately, whereas the validation loss reduced from 17.8% to a minimum of around 16.65%. Despite the overall pattern of both losses seeming to be reduced during the epochs, the validation loss began to fluctuate significantly after the 12th epoch, resulting in an early stop at the 15th epoch. The early stopping was necessarily executed so that the model does not memorize the training data.

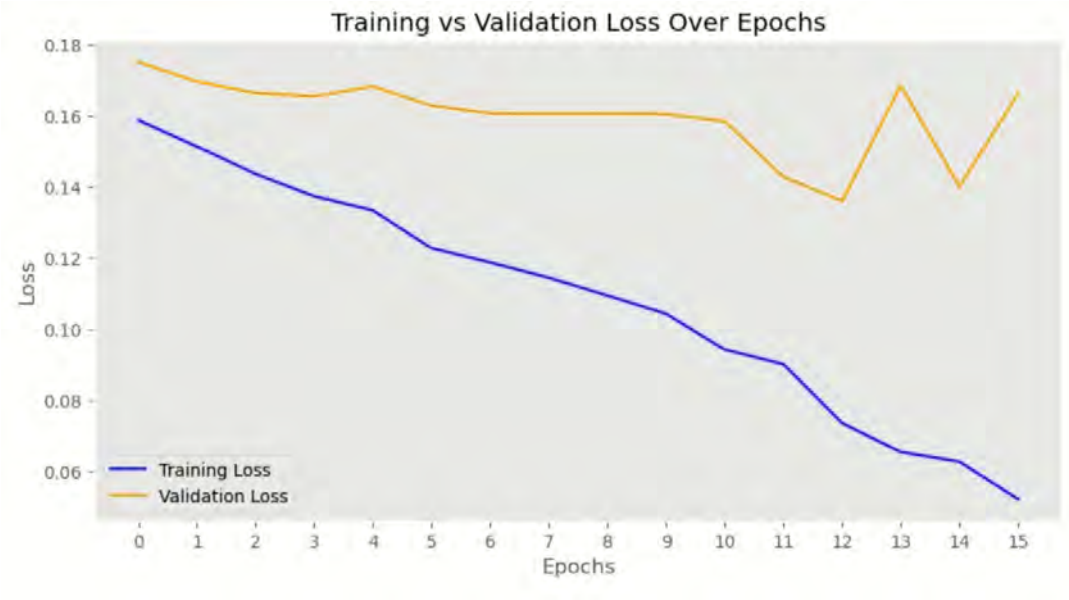

Figure 5.6: Training and Validation Loss of BiGRU Over Epochs

5.5 Combined Testing Evaluation and Analysis

5.5.1 Evaluation Metrics

The metric that we primarily focused on for classification was positive recall, as we deemed it more important to be able to identify as many anorexic users as possible. For early detection, we used the metric [2] that was proposed to be used in the competition:

$$ERDE_o(d,k) = \begin{cases} c_{fp} & \text{if } d = \text{positive AND ground truth} = \text{negative (FP)} \\ c_{fn} & \text{if } d = \text{negative AND ground truth} = \text{positive (FN)} \\ lc_o(k) \cdot c_{tp} & \text{if } d = \text{positive AND ground truth} = \text{positive (TP)} \\ 0 & \text{if } d = \text{negative AND ground truth} = \text{negative (TN)} \end{cases}$$

The ERDE (Early Risk Detection Error) metric evaluated both the accuracy of a binary decision and the delay incurred in making that decision. The delay was measured by counting the number of distinct submissions (e.g., posts or comments) that had been reviewed before reaching a conclusion. For example, if a user had made 90 submissions in total, with 9 submissions per chunk, and the system had provided a decision after analyzing the third chunk, the delay k would have been 27 submissions. Additionally, ERDE addressed the issue of imbalanced datasets, where negative cases significantly outnumbered positive ones. Consequently, the metric applied different weights to various types of errors. When a binary decision d had been made with a delay k, the prediction was classified into one of four categories: true positive (TP), true negative (TN), false positive (FP), or false negative (FN). The ERDE score was then calculated based on these classifications. The values for c_{fp} and c_{fn} were chosen based on the specific domain and the consequences of making false positive (FP) or false negative (FN) decisions. In system evaluations, c_{fn}

was kept constant at 1, while c_{fp} was adjusted based on the ratio of positive cases in the 2017 test dataset, ultimately being set to 0.1296. The factor $lc_o(k)$, which falls within the range of 0 to 1, was used to represent the cost associated with delayed identification of true positives. Additionally, c_{tp} was set equal to c_{fn} (with a value of 1) to reflect the severe implications of late detection, treating delayed discovery as essentially equivalent to missing the case entirely. The function $lc_o(k)$, a sigmoid curve, increased monotonically as the delay k grew. This penalty for latency was applied exclusively to true positives since late detection was not considered problematic for true negatives. True negatives, which are non-risk cases, do not require early identification, as there is no need for immediate action. The systems were therefore primarily focused on early identification of risk cases, while the detection of non-risk cases, regardless of timing, was simply a matter of ensuring these were correctly filtered out from the risk cases.

5.6 Performance Comparison With Previous Results

The reason we chose positive recall to be our main focus was to ensure the maximum number of accurately classified anorexia subjects since it is more important to identify these users so that they can be given proper treatment.

Team / Model	$ERDE_5$	$ERDE_{50}$	**Precision**	**Recall**	**F1 Score**
FHDO-BCSGD	12.15%	**5.96%**	0.75	**0.88**	0.81
FHDO-BCSGE	11.98%	6.61%	0.87	0.83	**0.85**
UNSLB	**11.40%**	7.82%	0.75	0.51	0.61
UNSLD	12.93%	9.85%	**0.91**	0.71	0.79
BERT+LSTM	13.79%	8.15%	0.35	0.93	0.51
BERT+BiLSTM	12.63%	6.99%	0.50	0.80	0.62
BERT+GRU	14.94%	8.67%	0.31	0.93	0.47
BERT+BiGRU	14.35%	8.41%	0.35	0.90	0.50

Table 5.1: Performance Comparison of Different Models

To evaluate the performance of our models, we compared our results with the previous ones obtained in the eRisk 2018 competition's Task2. We used the best results achieved in the competition for the comparison and bolded them as shown in Table 5.1. It can be seen that our best performing model in terms of positive class recall was achieved by two combinations: BERT+LSTM and BERT+GRU, with an astounding value of 0.93. In the competition, the leading positive recall and $ERDE_{50}$ of 0.88 and 5.96% respectively were both reached by the team FHDO-BCSGD, whereas FHDO-BCSGE, UNSLB and UNSLD were able to obtain the best F1-score (0.85), $ERDE_5$ (11.40%) and precision (0.91) respectively.

Our best recall value produced by both BERT+LSTM and BERT+GRU beat the top one achieved by FHDO-BCSGD. A high positive recall means that our two models were more effective at finding most of the true positive examples for the class. Recall measures the model's ability to correctly identify as many relevant instances

as possible. In detecting signs of anorexia from social media posts, our high recall score indicates that the models successfully captured the majority of the users who showed actual signs of having anorexia. This is especially important in cases like healthcare, where failing to detect a true positive could have significant implications. Nevertheless, increasing recall can lead to a rise in false positives, meaning models might incorrectly classify non-relevant instances as positive. This can lower precision, which measures how many of the predicted positives are actually correct. Since we aimed to maximize the recall using our models, we faced this problem where we attained our goal of identifying as many anorexic subjects as possible, but while doing so it also led to the models in falsely identifying many normal subjects as anorexic. Therefore, this had an adverse effect on our precision, where the highest positive class precision we were able to achieve was 0.50 with BERT+BiLSTM, and this ultimately compromised our F1-scores as well, indicating a lack of proper balance between identifying true positives and avoiding false positives. Our top value of F1-score was 0.62, which was again achieved by BERT+BiLSTM.

Considering all the above information it can be seen that among the models that we used, BERT+BiLSTM performed best with better overall results obtained in $ERDE_5$, $ERDE_{50}$, and positive class precision and F1-score. This is because BiLSTM layer improved the model's ability to capture extended relationships in the text. By processing the input sequence in both directions—forward and backward—it could better grasp the connections between earlier and later words, which aided in identifying patterns relevant to anorexia detection better and mitigate false positives. Nonetheless, the top positive recall performers BERT+LSTM and BERT+GRU need to be acknowledged as a high recall ensures that fewer positive cases were missed, which is crucial in areas where overlooking relevant instances can have serious consequences, even if it means accepting a higher rate of false positives.

Chapter 6

Limitations and Challenges

First and foremost, the dataset that we used was highly imbalanced, with only about 13% of the training data being from anorexia-positive users. Even though we tried to diminish the effects of this by using separate approaches like oversampling the positively labeled data, randomly mixing up the subjects during training, and ultimately finding the best technique to be the combination of shuffling up the training data and using weighted binary cross-entropy to penalize the labels inversely based on their frequency, the models that we used still failed to optimize the weights for minimum validation loss. Furthermore, irrespective of our models being able to recognize almost all of the anorexic users most of the time, there were instances where the models were unable to identify a few of these anorexic subjects that they would normally detect but were somehow able to properly classify the anorexia-negative users more accurately, resulting in less false positives. This was due to the random rearrangement of the training data at each epoch, for which we had to manually ensure the number of consecutive anorexic subjects in the training data to be at most two. However, this was not a dynamic solution and hence, we were not able to achieve the best results in terms of precision and F1-score.

Second, the negative instances of our training data had posts containing anorexia-related words for which our models had trouble classifying them as true negatives and ended up labeling them as positive. For example, a person who may not have anorexia may have posted content to increase awareness of the disorder, they may have commented on anorexia-related posts using phrases related to anorexia, and so on. All these led to the posts/comments of these users getting a high probability score for anorexia when in reality, they did not have the eating disorder.

Third, the detection threshold that we used was also hard-coded and not dynamic, as a person who showed signs of anorexia at an earlier chunk of their data may again be categorized as anorexic after more than the next three chunks of the data, but due to our rule being static, they may be misclassified. For early detection, we set a probability threshold of 70% which again, is just an assumption that a person crossing this threshold is without a question, anorexic. These may result in lower efficiency in distinguishing the positive class from the negative one.

Lastly, the models that we have trained may not be suitable in the long run because language keeps on having new additions and colloquial terms keep on changing with

time. As our models were trained on social media data where users use language that frequently alters, people with anorexia may express themselves differently further down the line. This will cause great difficulty for our models to capture the subtle nuances and will lead to poor performance in differentiating between anorexia and non-anorexia related posts.

Chapter 7

Conclusion

Due to the growing use of social media platforms, cyberbullying is on the rise, which has ultimately resulted in an increasing number of people being more conscious of their body shape and weight. The bullied users give subtle hints of their emotions in the text that they post on these online platforms and the need to analyze their texts to determine whether they are sufferers of various mental health disorders has become extremely important. Hence, our goal was to uncover a deep learning model to detect one of the many health disorders - anorexia, which would help identify whether a user was suffering from anorexia or not by analyzing their social media content. Moreover, limitations in datasets on the Internet for certain mental health disorders motivated us to evaluate our model from the perspective of early detection too, where we sought to correctly flag anorexic subjects in the shortest time possible, i.e. using as little data as we could.

7.1 Future Work

In the future, we plan to extend our current research by creating a dataset for binge eating disorder (BED) and applying our proposed approach of integrating BERT into deep learning models on it to analyze the performance and test the versatility of our approach. In this way, we can observe the performance on both ends of the spectrum - anorexia and BED. Furthermore, the dataset was highly imbalanced, and working with such a dataset can be very difficult. For this, we want to look for better methods to handle such imbalanced datasets. Moreover, the random rearrangement of the training data at each epoch was not a dynamic solution, limiting precision and F1 score, so we plan to find a dynamic solution for that as well. Next, the detection threshold and early detection criteria used were again hard-coded and not dynamic, so we can surely look for a better alternative for those too. Not just that, we also want to integrate explainable AI to understand the pattern so that the false positives can be reduced in a more efficient way. For example, SHAP or LIME explainable AI can be integrated to make the model more transparent, robust, high-performing, and more accurate, which is crucial for early risk detection tasks. These are some things that we are hoping to work on in the near future. Therefore, there is a great opportunity for the researchers to explore these aspects of our research.

Bibliography

[1] S. Cornegruta, R. Bakewell, S. Withey, and G. Montana, "Modelling radiological language with bidirectional long short-term memory networks," *arXiv preprint arXiv:1609.08409*, 2016.

[2] D. E. Losada and F. Crestani, "A test collection for research on depression and language use," in *International conference of the cross-language evaluation forum for European languages*, Springer, 2016, pp. 28–39.

[3] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[4] F. Sadeque, D. Xu, and S. Bethard, "Measuring the latency of depression detection in social media," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2018, pp. 495–503.

[5] Y.-T. Wang, H.-H. Huang, H.-H. Chen, and H. Chen, "A neural network approach to early risk detection of depression and anorexia on social media text.," in *CLEF (Working Notes)*, 2018, pp. 1–8.

[6] P. L. Úbeda, F. M. Plaza-del-Arco, M. C. Díaz-Galiano, L. A. U. Lopez, and M. T. Martín-Valdivia, "Detecting anorexia in spanish tweets," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 2019, pp. 655–663.

[7] S. G. Burdisso, M. Errecalde, and M. Montes-y-Gómez, "τ-ss3: A text classifier with dynamic n-grams for early risk detection over text streams," *Pattern Recognition Letters*, vol. 138, pp. 130–137, 2020.

[8] B. Funk, S. Sadeh-Sharvit, E. E. Fitzsimmons-Craft, *et al.*, "A framework for applying natural language processing in digital health interventions," *Journal of medical Internet research*, vol. 22, no. 2, e13855, 2020.

[9] Y. Hwang, H. J. Kim, H. J. Choi, and J. Lee, "Exploring abnormal behavior patterns of online users with emotional eating behavior: Topic modeling study," *Journal of Medical Internet Research*, vol. 22, no. 3, e15700, 2020.

[10] J. A. Benítez-Andrades, J. M. Alija-Pérez, I. García-Rodríguez, *et al.*, "Bert model-based approach for detecting categories of tweets in the field of eating disorders (ed)," in *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, IEEE, 2021, pp. 586–590.

[11] J. Chen, X. Huang, H. Jiang, and X. Miao, "Low-cost and device-free human activity recognition based on hierarchical learning model," *Sensors*, vol. 21, p. 2359, Mar. 2021. DOI: 10.3390/s21072359.

[12] A. Sammani, A. Bagheri, P. G. Heijden, *et al.*, "Automatic multilabel detection of icd10 codes in dutch cardiology discharge letters using neural networks," *npj Digital Medicine*, vol. 4, Apr. 2021. DOI: 10.1038/s41746-021-00404-9.

[13] A. S. Uban, B. Chulvi, and P. Rosso, "Understanding patterns of anorexia manifestations in social media data with deep learning," in *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, 2021, pp. 224–236.

[14] J. A. Benítez-Andrades, J.-M. Alija-Pérez, M.-E. Vidal, R. Pastor-Vargas, and M. T. García-Ordás, "Traditional machine learning models and bidirectional encoder representations from transformer (bert)–based automatic classification of tweets about eating disorders: Algorithm development and validation study," *JMIR medical informatics*, vol. 10, no. 2, e34492, 2022.

[15] Y. Hacohen-Kerner, N. Manor, M. Goldmeier, and E. Bachar, "Detection of anorexic girls-in blog posts written in hebrew using a combined heuristic ai and nlp method," *IEEE Access*, vol. 10, pp. 34 800–34 814, 2022.

[16] J. Sun, Y. Liu, J. Cui, and H. He, "Deep learning-based methods for natural hazard named entity recognition," *Scientific Reports*, vol. 12, p. 4598, Mar. 2022. DOI: 10.1038/s41598-022-08667-2.

[17] A. S. Uban, B. Chulvi, and P. Rosso, "Multi-aspect transfer learning for detecting low resource mental disorders on social media," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 3202–3219.

[18] Anishnama, *Understanding bidirectional lstm for sequential data processing*, en, May 2023. [Online]. Available: https://medium.com/@anishnama20/understanding-bidirectional-lstm-for-sequential-data-processing-b83d6283befc.

[19] Anishnama, *Understanding gated recurrent unit (gru) in deep learning*, en, May 2023. [Online]. Available: https://medium.com/@anishnama20/understanding-gated-recurrent-unit-gru-in-deep-learning-2e54923f3e2.

[20] R. Guerra, B. Pizarro, C. Aracena, *et al.*, "Cmm pln at mentalriskes: A traditional machine learning approach for detection of eating disorders and depression in chat messages," in *IberLEF (Working Notes). CEUR Workshop Proceedings*, 2023.

[21] M. Rujas, B. Merino-Barbancho, P. Arroyo, and G. Fico, "Development of a natural language processing-based system for characterizing eating disorders," in *IberLEF (Working Notes). CEUR Workshop Proceedings*, 2023.

[22] T. Vaj, *Gru vs. bi-gru: Which one is going to win?* en, Feb. 2023. [Online]. Available: https://vtiya.medium.com/gru-vs-bi-gru-which-one-is-going-to-win-58a45ede5fba.

[23] S. Dobilas, *Lstm recurrent neural networks — how to teach a network to remember the past*, en, Feb. 2024. [Online]. Available: https://towardsdatascience.com/lstm-recurrent-neural-networks-how-to-teach-a-network-to-remember-the-past-55e54c2ff22e.

[24] J. Parapar, P. Martín-Rodilla, D. E. Losada, and F. Crestani, *Clef erisk: Early risk prediction on the internet | clef 2024 workshop — erisk.irlab.org*, https://erisk.irlab.org/, [Accessed 20-05-2024], 2024.

[25] Anonymous, *Comparing bidirectional encoder representations from transformers (bert) with distilbert and bidirectional gated recurrent unit (bgru) for antisocial online behavior detection*, en-US, https://humboldt-wi.github.io/blog/research/information_systems_1920/bert_blog_post/, Accessed: 2024-10-10.

[26] Anonymous, *Eating Disorder Statistics | ANAD - National Association of Anorexia Nervosa and Associated Disorders — anad.org*, https://anad.org/eating-disorder-statistic/, [Accessed 22-01-2024].

[27] Anonymous, *Merge CSV files online into one file — merge-csv.com*, https://merge-csv.com/, [Accessed 20-05-2024].

[28] Anonymous, *What is an Eating Disorder: Types, Symptoms, Risks, and Causes — eatingdisorderhope.com*, https://www.eatingdisorderhope.com/information/eating-disorder, [Accessed 22-01-2024].

[29] Anonymous, *XML to CSV Converter | Online Tool — data.page*, https://data.page/xml/csv, [Accessed 20-05-2024].