Comparing Selective Masking Methods for Depression Detection in Social Media

Chanapa Pananookooln*
Asian Institute of Technology
Department of Information and
Communications Technologies
School of Engineering and Technology
chanapapan613@gmail.com

Jakrapop Akaranee
Asian Institute of Technology
Department of Information and
Communications Technologies
School of Engineering and Technology
jakrapop.a@gmail.com

Chaklam Silpasuwanchai
Asian Institute of Technology
Department of Information and
Communications Technologies
School of Engineering and Technology
chaklam@ait.ac.th

Identifying those at risk for depression is a crucial issue and social media provides an excellent platform for examining the linguistic patterns of depressed individuals. A significant challenge in depression classification problems is ensuring that prediction models are not overly dependent on topic keywords (i.e., depression keywords) such that it fails to predict when such keywords are unavailable. One promising approach is masking—that is, by selectively masking various words and asking the model to predict the masked words, the model is forced to learn the inherent language patterns of depression. This study evaluates seven masking techniques. Moreover, predicting the masked words during the pre-training or fine-tuning phase was also examined. Last, six class imbalanced ratios were compared to determine the robustness of masked words selection methods. Key findings demonstrate that selective masking outperforms random masking in terms of F1-score. The most accurate and robust models are identified. Our research also indicates

[*] Corresponding author.

Action Editor: Myle Ott. Submission received: 26 October 2022; revised version received: 12 February 2023; accepted for publication: 12 March 2023.

https://doi.org/10.1162/coli_a_00479

that reconstructing the masked words during the pre-training phase is more advantageous than during the fine-tuning phase. Further discussion and implications are discussed. This is the first study to comprehensively compare masked words selection methods, which has broad implications for the field of depression classification and general NLP. Our code can be found at: `https://github.com/chanapapan/Depression-Detection`.

1. Introduction

Depression is a growing problem, and it can lead to suicidal thoughts and mental disorders. Consequently, identifying individuals at risk remains an important problem. Because depressed people frequently use social media to seek assistance or to express their mental pain, social media is a valuable resource for studying the language associated with depression. In fact, numerous machine learning models have been proposed in an attempt to classify depression linguistically. Models such as SVM, log linear regression, decision tree, naive Bayes classifier, and dictionary learning have been used to classify handcrafted features such as count-based linguistic patterns, user profile, and user behavior (De Choudhury et al. 2013; Coppersmith, Dredze, and Harman 2014; Shen et al. 2017; Song et al. 2018; Zhang et al. 2021). Successively, end-to-end deep learning models incorporated with attention mechanisms have been used due to their superior performance and explainability (Sekulić and Strube 2020; Zogan et al. 2021; Zhang et al. 2021; Wołk, Chlasta, and Holas 2021; Lin et al. 2020). Recently, it was demonstrated that pre-trained models such as Bidirectional Encoder Representations from Transformers (BERT), which benefits from pre-training on a large dataset using self-supervised mode, outperformed other machine learning models on depression detection tasks (Zogan et al. 2021; Lin et al. 2020).

A challenging aspect of depression detection is preventing the model's excessive reliance on topic keywords, which renders it incapable of learning the inherent linguistic characteristics utilized by depressed users. We call this phenomenon the "keyword bias." For example, the model may overly rely on depression-related keywords due to its high probability of depression such that it fails to recognize more inherent language patterns of depression that may exhibit a weaker signal but remains one of the indicators of depression. Similar concerns were expressed by Moon et al. (2021), Yates, Cohan, and Goharian (2017), and Wolohan et al. (2018). For example, before classification, Yates, Cohan, and Goharian (2017) removed all posts made by depressed users that contain depression-related keywords, to prevent depressed users from being easily identified. They obtained an F1-score of 51%. Similarly, Wolohan et al. (2018) compared classification performance between all posts and posts excluding depression-related topics. When keywords were removed, the F1-score decreased by approximately 4 percent, from 73% to 68%. In any case, it is essential to note that Yates, Cohan, and Goharian (2017) only removed a small number of keywords, which were mostly names of mental health disorders, such as "depression" and "mdd," whereas many other keywords closely related to depression, such as "therapy" and "anxiety," remained. Moreover, Wolohan et al. (2018) only removed posts made in mental health subreddits, while other potential depression related keywords may remain present. Thus, there are many opportunities for further investigation. In addition, it has been found that depressed users tend to exhibit certain part-of-speech patterns or use more or less words in certain categories (Bucur, Podină, and Dinu 2021; Losada and Gamallo 2020; Bucur and Dinu 2020; De Choudhury et al. 2016; Zhang et al. 2021; De Choudhury et al.

2013; Coppersmith, Dredze, and Harman 2014; Stirman and Pennebaker 2001; Cohan et al. 2018; Morales, Scherer, and Levitan 2018). For example, the depressed group tends to use fewer proper nouns (Bucur and Dinu 2020) and more first person singular pronouns (De Choudhury et al. 2016) in comparison to the control group. Hence, it is important to train the depression detection model to capture these various patterns.

In the general natural language processing (NLP) field, a number of studies have been conducted to address the over-reliance of models on topic keywords. One promising approach is to mask various important words for models to learn language patterns from the context surrounding these masked words. Indeed, selective masking appears to be the most discussed approach. For example, Moon et al. (2021) proposed the use of the simple TF-IDF technique to identify the words to be masked based on word frequency. Aside from TF-IDF, Moon et al. (2021) also proposed the use of the average summation of attention scores across all samples to identify important words to be masked. Very similar to TF-IDF, Kawintiranon and Singh (2021) proposed the use of log-odds-ratio, which identified the words to be masked but based on word frequency variances across the whole corpus. Gu et al. (2020) proposed the use of a neural network, which identifies keywords based on how much a token, when added, increases the probability likelihood. Despite these advancements, these techniques have never been studied in the field of depression. In addition, no explicit comparison has been conducted between these masking techniques, even within the general NLP field.

This study comprehensively evaluates the efficacy of masked words selection methods on the Reddit Self-reported Depression Diagnosis (RSDD) dataset developed by Yates, Cohan, and Goharian (2017). A total of seven masked words selection methods (1 random and 6 selective) were compared: (1) random masking, (2) a depression lexicon (Losada and Gamallo 2020), (3) log-odds-ratio (Kawintiranon and Singh 2021), (4) TF-IDF (Moon et al. 2021), (5) summation of attention scores across all samples (Moon et al. 2021), (6) top attention scores across each sample, and, lastly, (7) Gu et al.'s neural network (Gu et al. 2020). In addition, two training methods were compared: the reconstruction of masked words either during the pre-training (Kawintiranon and Singh 2021) or the fine-tuning (Moon et al. 2021) phase. Last, six imbalanced class ratios (e.g., 1 user with depression versus 10 control users) were compared. Because imbalanced data is a common occurrence in depression datasets, it is essential to examine the robustness of masked words selection methods against varying imbalanced ratios. For example, in the original RSDD dataset, 12 control users were matched with 1 depressed user, which may cause the model to cheat by assigning all users to the control group to achieve good outcomes. The F1-score was the primary metric used. In summary, these are the primary research questions:

1. Do selective masking techniques that take keyword bias into account perform better than random masking?
2. If the answer is yes, which methods of selective masked words selection achieve the most accurate classification, and why? For example, how do dictionary-based methods (i.e., lexicons) compare to frequency-based methods (i.e., log-odds-ratio, TF-IDF) and neural network-based methods (i.e., attention)?
3. How do training methods affect the accuracy of classification? For instance, is it more advantageous to inject knowledge at an earlier stage (i.e., pre-training) or a later stage (i.e., fine-tuning)?

4. Which methods of masked words selection achieve the most robust performance on datasets with imbalance? It is intriguing to investigate which methods are most robust against extremely imbalanced data (e.g., a ratio of 1 depressed user to 10 control users).
5. Does any discernible pattern exist in the important words selected by the most effective masked words selection methods versus the less effective ones?

Based on F1-score, key findings include: (1) selective masking methods generally outperformed random masking, suggesting that selectively masking important words enhances a model's learning capacity; (2) at the extreme imbalanced ratio of 1:10, summation of attention was the top performer followed by top attention scores, while log-odds-ratio and TF-IDF were among the worst performers; (3) in all cases, the objective of reconstruction achieved the best performance when it was imposed during the pre-training phase; (4) summation of attention achieved the most robust performance across all imbalanced ratios; (5) the majority of masked words are words related to social, affective, cognitive, and biological processes. Further implications and future work are discussed.

The contributions of our study are as follows:

1. This study explicitly confirms the superiority of selective masking over random masking in the domain of depression classification.
2. This study comprehensively evaluates masked words selection methods, which poses implications on the keyword bias problem within the depression domain and beyond to the general NLP field.
3. This study compares training methods, that is, reconstruction of masked words either during the fine-tuning phase or the pre-training phase.
4. This study examines the robustness of masked words selection methods on different class imbalanced ratios, which poses implications for health- and crime-related fields where imbalanced data is prevalent.

2. Related Work

In this section, we review the methods of data collection, models used, and selective masking methods.

2.1 Data Collection

Developing a well-defined dataset for depression classification is more challenging than anticipated, and there are a number of potential failure points that may mislead us into believing that we achieve high accuracy. As a result, we made a comprehensive analysis of how previous researchers defined the dataset differently; this can act as a guideline for us to follow and help us avoid pitfalls. For better readability, we divide our analysis into subsubsections.

2.1.1 Manual vs. Automatic Labeling. It is necessary to label users into a control group and a depression group. The gold standard for identifying users in each group is to have them complete a standardized clinical depression survey. This strategy was utilized by De Choudhury et al. (2013), who invited crowdworkers to fill out two surveys: the CES-D (Center for Epidemiologic Studies Depression Scale) questionnaire and the Beck Depression Inventory (BDI). Users' Twitter profiles and self-reported depressive histories were also obtained. Despite producing a gold-standard dataset, this process is costly and time-consuming.

Subsequently, Coppersmith, Dredze, and Harman (2014) proposed an automatic data collection method. Their strategy targets Twitter users who have announced publicly that they have been diagnosed with a mental condition. Users of each mental disease category were acquired using a regular expression (i.e., "I was diagnosed with X."), and then further manually classified to determine whether the statement is authentic. For the control group, participants were picked at random from the overall Twitter user population. Some limitations were mentioned by the authors. First, this method can only capture the subset of users who discuss their condition openly. Second, the diagnoses' authenticity cannot be confirmed. Third, the control group may be contaminated with the presence of diagnosed users. Despite these limitations, the authors have proved the efficacy of their automatically created data by demonstrating that statistical classifiers can distinguish between users with four distinct mental health conditions within the dataset.

Following Coppersmith, Dredze, and Harman (2014), numerous studies improved upon the method (Coppersmith et al. 2015; Shen et al. 2017; Losada and Crestani 2016; Yates, Cohan, and Goharian 2017; Cohan et al. 2018; Zhang et al. 2021). For example, Shen et al. (2017) and Zhang et al. (2021) utilized additional regular expressions, such as (I was/ I am/ I've been) diagnosed with depression, and restricted non-depression users to those who had never submitted a tweet containing the word "depress." Yates, Cohan, and Goharian (2017) developed the RSDD dataset, which also has human annotators exclude false positive samples such as hypotheticals (e.g., "if I was diagnosed with depression"), negations (e.g., "it's not like I've been diagnosed with depression"), and quotes (e.g., "my brother announced 'I was just diagnosed with depression' "). For the control group, Yates, Cohan, and Goharian (2017) selected people who had never posted in a subreddit related to mental health and who had never used a phrase linked to depression or mental health. The RSDD dataset had 9,210 diagnosed users and 107,274 control users. The Self-reported Mental Health Diagnoses (SMHD) dataset (Cohan et al. 2018) expanded on RSDD by adding synonyms to matching patterns.

2.1.2 Time. The time period of each user to be included in the dataset is also a crucial factor to consider. Some studies simply cover the most recent posts of users (Coppersmith, Dredze, and Harman 2014; Coppersmith et al. 2015; Losada, Crestani, and Parapar 2018), whereas others considered the time of self-declared diagnosis. For instance, Shen et al. (2017) and Zhang et al. (2021) extracted posts from depression users one month and three months after the self-declared post, respectively. MacAvaney et al. (2018) evaluated diagnosis recency, which establishes when the diagnosis was made, and condition status, which indicates whether the diagnosed ailment is now active or has passed. In conclusion, we agree that the time of self-declared diagnoses should be considered such that the users' data should be collected after the users had declared the diagnoses, but not too long after, to ensure that the data we obtain is truly concomitant with the period during which the users had depression.

2.1.3 Removing Keywords. It is essential to prevent the model from overfitting to words associated with depression. Consequently, the model is susceptible to deception when depression keywords are not explicitly provided. Coppersmith et al. (2015) acknowledged this possibility but did not attempt to remove mentions of depression from the gathered dataset, whereas Yates, Cohan, and Goharian (2017) and Cohan et al. (2018) eliminated any postings by diagnosed users that fit either of the conditions, that is, that was posted in a mental health subreddit or included a mental disorder name. This issue was the focus of Wolohan et al. (2018) who compared classification performance under two conditions: one in which all user-generated content was included and one in which depression-related postings were omitted. The results demonstrated that the models' overall accuracy in detecting depression decreased by approximately 4 percent.

2.1.4 Imbalanced Class Ratio. Class imbalance is another concern in depression detection datasets. In the population of the real world, the control group significantly outnumbers the depression group. Whereas many studies maintained an equal proportion of the two groups (Coppersmith, Dredze, and Harman 2014; Zhang et al. 2021; AlSagri and Ykhlef 2020), others (De Choudhury et al. 2013; Shen et al. 2017; Coppersmith et al. 2015; Losada and Crestani 2016) addressed the imbalance in their dataset. In particular, the RSDD (Yates, Cohan, and Goharian 2017) and SMHD (Cohan et al. 2018) datasets matched each depressed user with 9 and 12 control users, respectively, to address the imbalance. Users with the smallest Hellinger distance between the diagnosed user's and the control user's subreddit post probability distributions were matched. This method, according to the authors, ensures that diagnosed users are paired with control users who are interested in similar subreddits and have similar activity levels, hence preventing biases based on the subreddits people participate in. We agree that an effective depression detection model must be robust on datasets with imbalances. For this reason, to evaluate the model's robustness, the performance of the model at various ratios between the control and depression groups must be evaluated.

2.1.5 Post vs User. The most typical strategy is to identify depression at the user level. However, a single user can provide an enormous amount of text data, which the model may not be able to process. Consequently, numerous solutions have been developed to address this issue. Some designed the model to process the data at the post level first then combine the features from all posts into a representation of the users (Sekulić and Strube 2020; Yates, Cohan, and Goharian 2017). In contrast, Jamil et al. (2017) demonstrated that a single tweet does not include sufficient information to detect whether a person is depressed. On the other hand, Zhang et al. (2021) created tweet chunks of 250 words by concatenating successive tweets from the same user and labeling them based on the user's label. The fact that they achieved identical F1-scores of depression detection on chunk-level and user-level data at 79% indicates that chunks of posts can also be categorized. Combining these findings with our premise that depressive language is present in all postings by depressed users, we conclude that chunking is an effective strategy.

2.2 Models

Previous studies trained machine learning models (e.g., SVM, log linear regression, decision tree, Naive Bayes classifier, dictionary learning, and neural networks) using features such as count-based linguistic patterns, user profile, and user behavior

(De Choudhury et al. 2013; Coppersmith, Dredze, and Harman 2014; Shen et al. 2017; Song et al. 2018; AlSagri and Ykhlef 2020; Zhang et al. 2021). Shen et al. (2017), for instance, extracted 6 depression-related feature groups (e.g., social network features, user profile characteristics, emotional features) and utilized them to train a multimodal dictionary learning model. On their own dataset, the resulting model was able to recognize users with depression with an F1-score of 84%. However, these features are handcrafted, which involves considerable work.

Using sliding *n*-gram windows, a convolutional neural network (CNN) can be applied to a text sequence. Consequently, this was utilized in a number of investigations (Yates, Cohan, and Goharian 2017; Orabi et al. 2018; Rao et al. 2020). A CNN user-model was suggested by Yates, Cohan, and Goharian (2017). It processes each user's postings and merges them to generate a vector representation of the user's activity, which was then fed to the classification layers. The F1-score on the RSDD dataset was 51%. Rao et al. (2020) enhanced Yates, Cohan, and Goharian's (2017) model by incorporating gating weights. Models incorporating an attention mechanism have also been applied more widely (Sekulić and Strube 2020; Zogan et al. 2021; Zhang et al. 2021; Wołk, Chlasta, and Holas 2021; Lin et al. 2020). On the SMHD dataset, Sekulić and Strube (2020) utilized a hierarchical attention network and produced an F1-score of 68.28%, surpassing logistic regression, SVM, and Supervised FastText by more than 10%. They also assessed attention weights on a word-by-word basis and compared the most attended terms to a previous depression study. Personal pronouns are crucial in identifying depressed authors from non-depressed authors, as demonstrated by the results. This highlighted the advantages of the attention mechanism in interpretability.

Recent research has used transformer-based pre-trained models like BERT and XLNet (Zogan et al. 2021; Lin et al. 2020). Zogan et al. (2021) proposed a hybrid framework consisting of a user behavior network and a posting history–aware network that utilized BERT and BART, achieving an F1-score of 91.2% on Shen et al.'s (2017) dataset. Similarly, Zhang et al. (2021) utilized XLNet, which outperformed SVM by more than 5% on their largest Twitter depression dataset. These results demonstrated the promising potential of a pre-trained transformer-based model for depression detection.

2.3 Selective Masking in Masked Language Model

It has been demonstrated that a pre-trained language model (e.g., BERT [Devlin et al. 2018]) is useful for improving several NLP problems. To further enhance BERT, it is possible to pre-train the model with domain-specific information. For example, Sun et al. (2019) further pre-trains BERT using in-domain data before fine-tuning BERT for the target task. The proposed approach achieved a new state-of-the-art on 8 text categorization datasets. The gain in performance from further pre-training were also proven in other domains such as biomedicine, computer science, and law (Chalkidis et al. 2020; Gu et al. 2021; Gururangan et al. 2020).

Selective masking methods (rather than random masking) may also be utilized to improve BERT in order to inject knowledge into the model. Kawintiranon and Singh's (2021) contribution to the selective masking method on the stance detection task was noted. For identifying the most distinguishing stance terms, they proposed log-odds-ratio using Dirichlet. The significant tokens were then masked from the unlabeled election data during additional pre-training of BERT, and the model was pre-trained on the masked language modeling task to recover the significant words. The outcomes demonstrated that their knowledge-enhanced model outperformed the original BERT by 3–5%. Another study on sentiment analysis by Tian et al. (2020) presented a similar approach in which they constructed a distorted version of the input sequences by removing the sentiment information and then required the transformer to recover the deleted information. This method intended to incorporate word-, polarity-, and aspect-level sentiment information into a previously trained sentiment representation. The outcomes considerably exceeded traditional RoBERTa and established a new state-of-the-art for sentiment analysis on numerous datasets. Due to the fact that many clinical NLP tasks are oriented on entities, Lin et al. (2021) presented an entity-centric masking method to include domain knowledge into the model. Their technique performed exceptionally well on three clinical NLP tests. Gu et al. (2020) proposed a selective masking approach that measures the significance of each token in sequences; that is, they trained a neural network to learn the implicit token-selection criteria. The neural network is then used to select tokens to mask from the unsupervised dataset within the domain. The findings of the experiment indicated that their selective masking strategy consistently outperformed the random masking method. Moon et al. (2021) proposed masked keyword regularization (MASKER) with two regularization techniques, where the first one concentrates on reconstructing the masked words and the second one compels the model to generate low-confidence predictions when all context words except the keyword are masked. It enhanced out-of-distribution detection and cross-domain generalization without compromising classification accuracy.

In general, selective masking is preferable to random masking. We hypothesize that this strategy would be highly advantageous for the depression detection challenge, in which we seek to promote the model to learn depression-related language from users' social media posts, rather than depending mainly on depression-related keywords.

3. Methodology

Figures 1 and 2 illustrate the overall methodology. Seven masked words selection methods, two training methods (reconstruction of masked words during pre-training versus fine-tuning), and six class imbalanced ratios were compared. Notably, we have two datasets: one for additional pre-training (i.e., in-domain dataset) and one for the classification task itself (i.e., classification datasets). Because the in-domain dataset was used to further pre-train BERT in a self-supervised manner, it does not undergo the

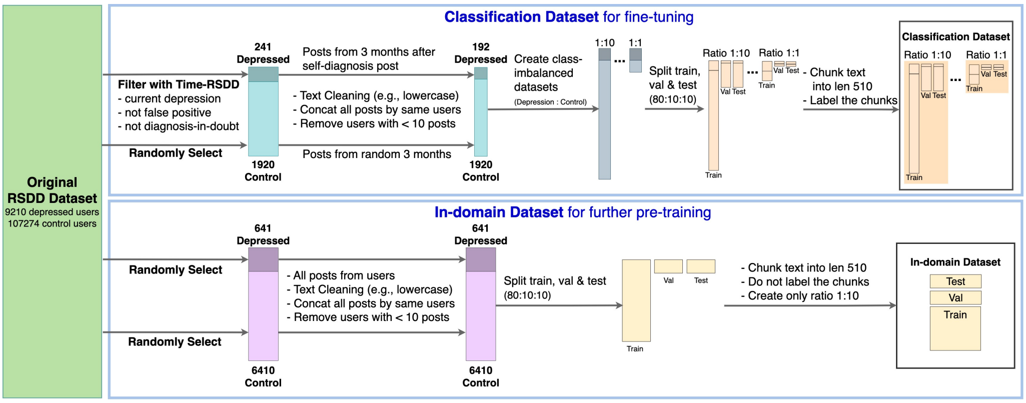
Original
RSDD Dataset
9210 depressed users
107274 control users
Classification Dataset for fine-tuning
Filter with Time-RSDD
- current depression
- not false positive
- not diagnosis-in-doubt
Randomly Select
241
Depressed
1920
Control
Posts from 3 months after self-diagnosis post
- Text Cleaning (e.g., lowercase)
- Concat all posts by same users
- Remove users with < 10 posts
Posts from random 3 months
192
Depressed
1920
Control
Create class-imbalanced datasets
(Depression : Control)
1:10
1:1
Split train, val & test (80:10:10)
Ratio 1:10
Ratio 1:1
Train
Val
Test
- Chunk text into len 510
- Label the chunks
Classification Dataset
Ratio 1:10
Ratio 1:1
Train
Val
Test
In-domain Dataset for further pre-training
Randomly Select
Randomly Select
641
Depressed
6410
Control
- All posts from users
- Text Cleaning (e.g., lowercase)
- Concat all posts by same users
- Remove users with < 10 posts
641
Depressed
6410
Control
Split train, val & test (80:10:10)
Train
Val
Test
- Chunk text into len 510
- Do not label the chunks
- Create only ratio 1:10
In-domain Dataset
Test
Val
Train

Figure 1
Data preprocessing and dataset formation. len = length.

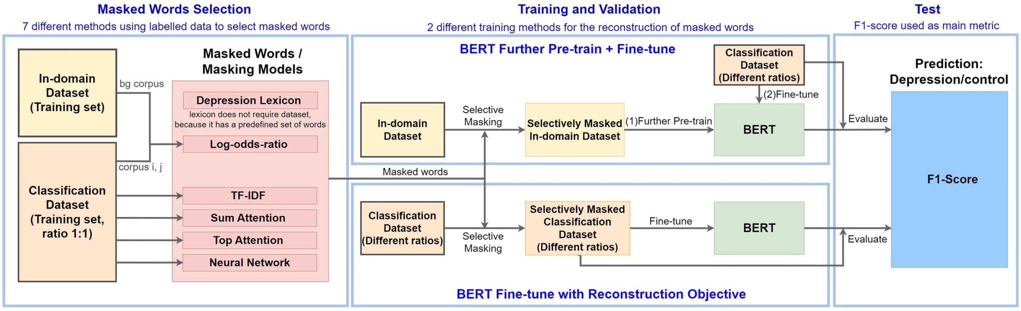
Masked Words Selection
7 different methods using labelled data to select masked words
In-domain Dataset (Training set)
bg corpus
Masked Words / Masking Models
Depression Lexicon
lexicon does not require dataset, because it has a predefined set of words
Log-odds-ratio
corpus i, j
Classification Dataset (Training set, ratio 1:1)
TF-IDF
Sum Attention
Top Attention
Neural Network
Training and Validation
2 different training methods for the reconstruction of masked words
BERT Further Pre-train + Fine-tune
Classification Dataset (Different ratios)
(2)Fine-tune
In-domain Dataset
Selective Masking
Selectively Masked In-domain Dataset
(1)Further Pre-train
BERT
Masked words
Classification Dataset (Different ratios)
Selective Masking
Selectively Masked Classification Dataset (Different ratios)
Fine-tune
BERT
BERT Fine-tune with Reconstruction Objective
Test
F1-score used as main metric
Prediction: Depression/control
Evaluate
F1-Score
Evaluate

Figure 2
Masked words selection, training, validation, and testing.

same preprocessing step as the classification tasks, for example, labeling is not necessary for the in-domain dataset. After creating the two datasets, we extracted masked words using our seven methods. Here, a classification dataset with a ratio of 1:1 was used to learn the masked words, in order to avoid causing any class bias in the masked word lists. Once these masked words were extracted, two training methods were compared: either to further pre-train BERT by masking the selected masked words in the in-domain dataset, or to simply proceed to fine-tune the classification dataset while adding an additional objective to recover the masked words. The resulting BERT was subsequently evaluated on the test set. During this phase of training, validation, and testing, different classification dataset ratios were compared. For further details, we describe the dataset, masked words selection methods, training methods, and metrics.

3.1 Dataset

The RSDD dataset (Yates, Cohan, and Goharian 2017) was chosen for a variety of reasons. First, high-precision matching patterns and human annotation were used to eliminate self-declared posts with false positives. Second, the RSDD-Time dataset (MacAvaney et al. 2018) contains the temporal information of some users from the RSDD dataset, which is crucial for selecting data that corresponds to the period in which the users were depressed. Lastly, the RSDD dataset is one of the largest depression detection datasets where deep learning models would be advantageous.

There are 9,210 depressed users and 107,274 control users in the RSDD dataset. The average number of posts per user is 969 (median: 646), and the average post length is 148 tokens (median 74).

On the basis of this dataset, we created two datasets tailored to our needs: an in-domain dataset to further pre-train BERT and a classification dataset for the actual task. We describe the two datasets in greater detail below.

3.1.1 Classification Dataset. We began by identifying depressed users who met our criteria. We used the temporal information in the Time-RSDD dataset to select only depressed users who were identified by the dataset as currently having the condition, not in the false positive or diagnosis-in-doubt groups from the RSDD dataset. Of the 9,210 depressed users in the RSDD dataset, 241 met the inclusion criteria.

Then, similar to Zhang et al. (2021), we collected all posts made by each depressed user in the 3 months following their self-diagnosis post. Then, we removed URLs,

Table 1
Number of users and chunks in the classification and in-domain dataset.

Control: Depression	# Control Users	# Depression Users	Total Chunks
Classification Dataset			
1:1	192	192	5,752
2:1	384	192	7,410
4:1	768	192	10,573
6:1	1,152	192	13,479
8:1	1,536	192	16,565
10:1	1,920	192	19,877
In-domain Dataset			
10:1	6,140	614	400,303

hashtags, emojis, smileys, punctuation (except for . ′ ! ? $ % &), unescaped HTML tags, and lowercased the text.

Users with fewer than 10 posts were removed (since according to the mean and median of the post length we needed to make sure we have enough data to make up at least one chunk). We were thus left with 192 depressed users (from 241 users). All posts of each user were concatenated into one long sequence, resulting in 192 text sequences.

To replicate the ratio of depression-to-control users in previous research (Zhang et al. 2021; De Choudhury et al. 2013; Cohan et al. 2018; Shen et al. 2017; Coppersmith et al. 2015; Losada and Crestani 2016), six classification datasets with ratios of 1:1, 1:2, 1:4, 1:6, 1:8, and 1:10 were created. The dataset with different class imbalanced ratios was compiled by collecting a larger amount of control users than depressed users. First, to collect control users for the dataset with the highest depression-to-control users ratio of 1:10, 1,920 (192 × 10), RSDD control users were selected at random. Because there were no self-diagnosis posts in the control group, we randomly selected 3 months′ worth of posts from each user. The text was then cleansed and concatenated in the same manner as the depressed group, resulting in 1,920 text sequences. Combining a subset of the 1,920 control text sequences with the 192 depression text sequences (i.e., 192:192, 192:384, $\cdots$, 192:1920) yielded datasets with varying class imbalanced ratios. For all datasets, the sequences were divided in ratio of 80:10:10 into train, validation, and test sets, respectively. Across all experiments, the same seed was used. Table 1 summarizes the datasets.

To accommodate BERT′s maximum input length (i.e., 512), we tokenized all sequences using BERTtokenizer and then divided each tokenized sequence into chunks of size 510 (512 including the [CLS] and [SEP] tokens). We tokenized the text prior to chunking because BERT uses the WordPiece tokenizer, which can divide words into subwords. Any left over chunks with length of less than 510 were removed. The chunks were then labeled using the user′s label.

3.1.2 In-domain Dataset. We developed the in-domain dataset to pre-train BERT in a self-supervised manner. To reduce computational time, only a subset of the RSDD was used. Twenty percent of depressed users (614 users) were selected at random from the RSDD dataset. Then, to match the most unbalanced ratio (1:10) of our classification dataset, 6,140 control users were randomly selected from the RSDD dataset. Each user′s posts were collected, cleaned, and concatenated in the same manner as the classification dataset, yielding a total of 6,754 sequences. Then, we split these sequences into train, validation, and test set with a ratio of 80:10:10, respectively, before tokenizing and chunking them into length 510.

3.2 Masking Word Selection

We compared seven different masking word selection methods that select and mask the most important or significant words from the depression and control corpus in order for the language model to learn the context or linguistic patterns of depression and control users that benefit depression detection. The seven techniques consisted of random masking, depression lexicon (Losada and Gamallo 2020), log-odds-ratio (Kawintiranon and Singh 2021), TF-IDF, summation of attention scores across all samples (Moon et al. 2021), neural network (Gu et al. 2020), and top attention scores across each sample.

Note that, using the depression lexicon, log-odds-ratio, TF-IDF, and the sum of attention scores across all samples, we obtained specific sets of words or tokens that were the same across the entire dataset. The size of masked word set for each of these four methods was empirically set to $k \times 2$ where k was empirically set to 1,500. This was set such that the percentage of masked words in each sequence was not too much larger or smaller than 15% (76 tokens), which is the standard number of masked tokens in pre-training of standard BERT (Devlin et al. 2018). However, for neural network and top attention scores across each sample, a model was used to specify which tokens or words would be masked, which can vary from sequence to sequence. Hence, the number of masked words in each sequence can be different. Thus, we refer in Figure 2 to Masked Words / Masking Models.

We used our classification dataset with a ratio of 1:1 as the input for all masked words selection methods. This is to avoid causing any class bias in the masked word lists that could be transferred to the model. Thus, if we were to use the trained model to classify real-world samples, the model would not be biased by the masked words, which is beneficial since we do not know the class imbalanced ratio of the distribution that the samples came from. Moreover, as the masked words were obtained from a balanced dataset, we would be able to apply these masked words to mask and train the model on datasets with any ratio of class imbalance. This is essential in a real-world application where the ratio of control to depression samples may not be constant. Additionally, only the training set was utilized to prevent data leakage to the validation and test sets.

3.2.1 Random. Since whole-word masking has been adopted as the standard approach because it forces the language model to capture more contextual semantic dependencies (Gu et al. 2021), we implemented whole-word masking BERT[1] (WW-BERT) to randomly select words until 15% of the tokens are selected in each sequence.

3.2.2 Depression Lexicon. A pre-existing lexicon is the most convenient way to identify words that are significant to depression. Therefore, we used the best-performing depression specific lexica for detecting signs of depression from Losada and Gamallo (2020), which improved De Choudhury et al.'s (2013) lexica by obtaining non-ambiguous adjectives and expanding the WordNet. This lexicon included 112 words. Some word examples from the lexicon are "anxiety," "drugs," "attacks," "antidepressant," and "psychotherapy."

[1] `https://github.com/google-research/bert`.

3.2.3 Log-odds-ratio. Frequency is one of the most prevalent methods for extracting key words from a corpus. Log-odds-ratio is a frequency-based approach proposed by Kawintiranon and Singh (2021) to compute significant words for two corpora by taking into account the variance in a word's frequency and using word frequencies from a background corpus to reduce the noise caused by rare words.

The usage difference for word w among two corpora was computed as shown in Equation (1), where n^i and n^j is the size of corpus i and j. y^i_w and y^j_w represent the word count of w in corpus i and j, respectively. α_0 is the size of the background corpus and α_w is the word count of w in the background corpus.

$$\delta_w^{i-j} = \log \frac{y^i_w + \alpha_w}{n^i + \alpha_0 - y^i_w - \alpha_w} - \log \frac{y^j_w + \alpha_w}{n^j + \alpha_0 - y^j_w - \alpha_w} \tag{1}$$

To measure the significance of each word, the variance (σ^2) of log-odds-ratio is computed using Equation (2), then the Z-score is computed using Equation (3). A higher Z-score means that the word w is more important in corpus i than in corpus j. In the case of a low score, the opposite is true.

$$\sigma^2(\delta_w^{(i-j)}) \approx \frac{1}{y^i_w + \alpha_w} + \frac{1}{y^j_w + \alpha_w} \tag{2}$$

$$Z = \frac{\delta_w^{(i-j)}}{\sqrt{\sigma^2(\delta_w^{(i-j)})}} \tag{3}$$

We used all text from the depression class of our classification dataset as corpus i and all text from the control class as corpus j. Moreover, because a background corpus was needed in this algorithm, we used all the text from the training set of our in-domain dataset as the background corpus. Due to the fact that this is a frequency-based approach, punctuation and stopwords that appear very frequently may affect the output. Hence, we removed all punctuation (except ′ which were left for decontraction) and stop words from the text before passing it to the algorithm. Then, according to Kawintiranon and Singh (2021), we collected the resulting top and bottom k words, yielding a list of $k \times 2$ masked words.

3.2.4 TF-IDF. TF-IDF is another frequency-based method that indicates the significance of a word in a corpus document. Due to its frequency-based nature, we cleaned the text by removing punctuation and stopwords in the same manner as log-odds-ratio. Initially, we divided the input text into depression corpus and control corpus and identified important words as tokens with the highest TF-IDF scores from each class, using the method proposed in Moon et al. (2021). However, when the top k words were determined, the majority of them were shared by both classes. As a result, in order to mitigate the redundancy issue, we modified the methodology by obtaining the 5,477 words with the highest TF-IDF scores from each corpus. Then, we obtained $k \times 2$ masked words consisting of k words from top depression words that were not in top control words and another k words from top control words that were not in top depression words.

3.2.5 Summation of Attention Score. Attention is another promising method for measuring the significance of words. First, we fine-tuned BERT on our classification dataset. Then, we compiled the attention scores for each token across all correctly classified samples and normalized them by token frequency. Unlike the method in Moon et al. (2021), where attention scores were collected across all samples, our method only considered the attention scores that contributed to accurate predictions. We also observed that extremely rare words influence the attention scores, so we set the attention scores to 0 for tokens with a frequency of less than 10. Finally, the $k \times 2$ tokens with the highest sum of attention scores were chosen.

The attention score was calculated as follows. Let $\mathbf{a} = [a_1, \cdots, a_T] \in \mathbb{R}^T$ be attention values of the document embedding, where a_i corresponds to input token t_i. Then, the attention-based score of token t is computed by

$$s^{\text{attn}}(t) = \sum_{(x,y)\in\mathcal{D}} \frac{1}{n_{t,x}} \sum_{i\in\{1,\cdots,T\}} \mathbb{I}(t_i = t) \cdot \frac{a_i}{\|\mathbf{a}\|} \tag{4}$$

where $\mathbb{I}$ is an indicator function and $\|\cdot\|$ is ℓ_2-norm.

3.2.6 Neural Network. Next, we apply the method of Gu et al. (2020), who proposed training a neural network for important words selection. First, we fine-tuned BERT on our classification dataset. Then, the important tokens were selected according to the following steps.

Given the n-token input sequence $\mathbf{s} = (w_1, w_2, \cdots, w_n)$, a sequence buffer ($\mathbf{s}'$) is used to evaluate the tokens one by one. At time step 0, $\mathbf{s}'$ is empty. Then, each token w_i is sequentially added to $\mathbf{s}'$ and the task-specific score of w_i, denoted by $\mathbf{S}(w_i)$, is calculated. If the score is lower than threshold δ, w_i is considered an important token.

The token w_i's score is the difference of probability likelihood between the original input sequence $\mathbf{s}$ and the buffer after adding w_i, which is denoted by $\mathbf{s}'_{i-1}w_i$:

$$\mathbf{S}(w_i) = P(y_t|\mathbf{s}) - P(y_t|\mathbf{s}'_{t-1}w_i) \tag{5}$$

where y_t is the ground truth label of the input $\mathbf{s}$ and $P(y_t|*)$ is the probability likelihood computed by the BERT model trained on the classification task. We empirically set δ to 0.001. The important token criterion is $S(w_i) < \delta$, which means that after adding w_i, the fine-tuned BERT model can correctly classify the incomplete sequence buffer with a probability likelihood close to the complete sequence. If w_i was considered important it would be removed from the buffer before adding the next token. After the selection, important words were annotated "1" and others, "0."

Next, another BERT model was trained on the token classification task to learn the token selection rule. The token classification model was trained with a learning rate = 5e–7 (we tested 5e–6 and 5e–7, and 5e–7 gained better performance) for 100 epochs, and the best model was selected by validation loss. This token selection BERT model was then used to classify each token in the dataset that we wanted to selectively mask. If the classification result is "1," then the token will be regarded as important and will be masked.

3.2.7 Top Attention Score. Here, we implemented another method for masked words selection, adapted from Moon et al. (2021) and Gu et al. (2020), in which we hypothesized that the most influential words are those with the highest attention scores, and that the most influential words can vary depending on the context, so no specific list of words was required. First, we fine-tuned BERT on our classification dataset. Next, we identified the most important words by using BERT's attention scores to select the top words with the highest attention scores (if the word consisted of many tokens, we averaged the score across all tokens of that word), which made up 15% of tokens from each sample. Then, another BERT would be trained on a token classification task to learn this token selection rule similar to the method in Gu et al. (2020). The token classification model was trained with a learning rate = 5e–6 for 100 epochs, and the best model was selected by validation loss. Then, in the same way as with the neural network method, we used the model to specify which tokens should be masked.

3.3 Training Methods

After identifying the masked words, we trained BERT to recover the selectively masked words from the dataset so as to inject depression detection knowledge into the model. We tested two different training methods, one in which the knowledge is injected during the pre-training phase and the other during the fine-tuning phase. In this section, we describe the two training methods, the process of selectively masking the datasets for each method, and the standard BERT fine-tune model that served as our baseline. All models' skeletons were loaded with BERT-base-uncased[2] weights.

3.3.1 Standard BERT Fine-tune. Adding a classification head and fine-tuning it on the classification dataset is the simplest way to fine-tune BERT for a classification task. As a baseline model, we train a standard BERT fine-tuned model using BERT-base-uncased weights to fine-tune with a cross-entropy objective on our classification dataset without any pre-training or additional training objectives. Training was performed six times, once on each of our classification datasets, resulting in six models for six distinct class imbalanced datasets.

The fine-tuning hyperparameters were as follows: learning rate = 5e–7 and batch size = 32. The models were further pre-trained for 50 epochs and the best models were saved according to the validation loss.

3.3.2 BERT Further Pre-train + Fine-tune. The first method involves training BERT on a selectively masked in-domain dataset during the BERT further pre-training step and then fine-tuning BERT for the downstream task.

We used masked words/masking models to identify all significant words in each sequence. If the number of selected tokens was less than or greater than 15%, we shuffled the identified significance words then randomly selected or deselected words until 15% of tokens were selected. This method maintained a constant number of masked words across sequences. To imitate the standard BERT pre-training, we randomly replaced 80% of the selected words with mask tokens, 10% with random tokens, and 10% with the original tokens for all masking methods. Note that the next sentence prediction task was not included in further pre-training.

The further pre-training hyperparameters were as follows: learning rate = 1e–4 and batch size = 16. The models were further pre-trained for 6 epochs and the best models were saved according to the validation loss.

[2] https://huggingface.co/bert-base-uncased.

Finally, for each masking method, we used the weights from further pre-training to perform BERT fine-tuning 6 different times, once on each of our classification dataset. This results in 6 models per masking method. The hyperparameters for fine-tuning were as follows: learning rate = 5e–7 (we experimented with learning rates of 1e–5, 5e–6, 1e–5, 5e–7, but 5e–7 allowed for the smoothest convergence and was chosen), batch size = 32. The models were refined for 50 epochs, and the models with the lowest validation loss were saved.

3.3.3 BERT Fine-tune with Reconstruction Objective. Another method to inject knowledge was to let the model recover masked words as one of the fine-tuning objectives. We used the MASKER model proposed by Moon et al. (2021) with the aim to reduce over-reliance issues and force the model to learn the context of important words. In their model, there was no further pre-training of BERT, however, two regularization techniques were added in BERT fine-tuning.

The first regularization was masked keyword reconstruction (MKR) where they forced the model to look at the surrounding context by guiding the model to reconstruct the words from masked documents.

Let $\tilde{\mathbf{k}}$ be a random subset of the full keywords set $\mathbf{k}$. Each element of $\tilde{\mathbf{k}}$ was chosen with probability p independently. $\tilde{\mathbf{k}}$ were masked from the original document ($\mathbf{x}$) and the masked document is denoted as $\tilde{\mathbf{x}} = \mathbf{x} - \tilde{\mathbf{k}}$. Consequently, the MKR loss is

$$\mathcal{L}_{MKR}(\tilde{\mathbf{x}}, v) := \sum_{i \in \mathrm{index}(\tilde{\mathbf{k}})} \mathcal{L}_{CE}(f_{MKR}(\tilde{\mathbf{x}})_i, v_i) \tag{6}$$

where index($\tilde{\mathbf{k}}$) is the index of keywords $\tilde{\mathbf{k}}$ in the original document $\mathbf{x}$. f_{MKR} represents the MKR part of the model which takes $\tilde{\mathbf{x}}$ as input and v_i is the index of the keywords with respect to the vocabulary set.

The second regularization was masked entropy regularization (MER). This technique forced the model to look at the context of the sequence (not the keywords) to make correct predictions. It does so by regularizing the model to produce low-confidence prediction when all the words except the important words were masked.

Let $\hat{\mathbf{c}}$ be a randomly chosen subset of the full context words denoted as $\mathbf{c} = \mathbf{x} - \mathbf{k}$, where each element is chosen with probability q independently. We mask $\hat{\mathbf{c}}$ from the original document ($\mathbf{x}$) and get the context-masked document $\hat{\mathbf{x}} = \mathbf{x} - \hat{\mathbf{c}}$. Then, the MER loss is computed as

$$\mathcal{L}_{MER}(\hat{\mathbf{x}}) := D_{KL}(\mathcal{U}(y) \| f_{MER}(\hat{\mathbf{x}})) \tag{7}$$

where D_{KL} is the KL-divergence and $\mathcal{U}(y)$ is a uniform distribution and f_{MER} represents the MER part of the model that takes $\hat{\mathbf{x}}$ as input.

BERT-base-uncased weights were loaded to the backbone of the MASKER model. The final objective function for MASKER was a standard cross entropy loss for sequence prediction (depression/control), MKR loss, and MER loss. To sum up, the final objective is given by

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{MKR}}\mathcal{L}_{\mathrm{MKR}} + \lambda_{\mathrm{MER}}\mathcal{L}_{\mathrm{MER}} \tag{8}$$

where $\mathcal{L}_{MKR}$ and $\mathcal{L}_{MER}$ are hyperparameters for the MKR and MER losses, respectively. λ_{MKR} and λ_{MER} of 0.001 were used (Moon et al. 2021).

To imitate the random masking in standard BERT, for MKR we randomly masked words until 15% of the tokens were masked and for MER we masked all non-selected tokens.

For other masking methods, according to Moon et al. (2021), each masked word in a sequence was masked with probability p independently in MKR, while each non-masked word in a sequence was masked with probability q independently in MER. We set p and q to 0.9 to make sure that sufficient words were masked in each sequence.

For each masking method, training was performed 6 different times, once on each of our classification datasets, which results in 6 models per masking method. The hyperparameters for training all MASKER models were as follows: learning rate = 1e–6 (we experimented with learning rate of 2e–6 and 1e–6, but 1e–6 achieved faster convergence) and batch size = 8. The models were trained for 50 epochs and the best models were saved according to the F1-score.

In order to examine the effects of each regularization technique, an ablation study was conducted where only MKR loss or only MER loss was added to the objective function, namely, BERT fine-tune with MKR and BERT fine-tune with MER. The final objective is given by

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{MKR}}\mathcal{L}_{\text{MKR}} \tag{9}$$

and

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{MER}}\mathcal{L}_{\text{MER}} \tag{10}$$

respectively. The same hyperparameters as in BERT fine-tune with reconstruction objective were used.

3.4 Metrics

We assess the depression detection performance of each model using a test set with the same imbalanced ratio and masking method as the dataset used for training. The performance of the final classification models were evaluated in terms of F1-scores since it is a metric that takes class imbalance into account.

F1-score is computed as the following:

$$F1 = \frac{2 \times (precision \times recall)}{(precision + recall)} \tag{11}$$

4. Experimental Results

Tables 2–5 and Figures 3–6 summarize the F1-scores across our comparisons. We present the results according to each factor, that is, masking methods, training methods, class imbalanced ratios, and masked words.

4.1 Masking Methods

Considering the BERT further pre-train + fine-tune training approach (see Table 2), the average of all ratios revealed that sum attention outperformed other methods, followed by top attention and neural network, respectively. The worst performer was

Table 2
F1-scores of BERT further pre-train + fine-tune trained on different imbalanced datasets with different masking methods. Highest score and scores no less than one percent of the highest across each class imbalanced ratio are **bolded**.

Control: Depression	Standard BERT fine-tune	BERT further pre-train + fine-tune							AVG ± SD
		Random	Lexicon	Log-odds	TF-IDF	Sum Att	Top Att	NN	
1	78.05	**78.74**	77.78	78.16	77.99	**79.25**	**78.52**	78.01	**78.35± 0.54**
2	71.10	**72.44**	71.50	**72.17**	**72.03**	**72.93**	71.45	**72.20**	72.10± 0.56
4	60.03	60.36	**63.29**	62.01	61.63	**63.65**	**64.17**	61.15	62.32± 1.44
6	54.83	51.72	55.45	**59.80**	54.99	**59.53**	55.78	56.18	56.21± 3.04
8	55.00	50.70	49.40	50.18	53.28	**59.94**	55.84	53.96	53.33± 4.05
10	52.20	50.46	49.23	46.63	47.42	**56.20**	53.06	52.11	50.73± 3.61
AVG ± SD	61.87 ± 10.40	60.74 ± 12.24	61.11 ± 11.84	61.49 ± 12.20	61.22 ± 11.75	**65.25 ± 10.56**	63.14 ± 8.95	62.27 ± 10.16	

Table 3
F1-scores of BERT fine-tune with additional reconstruction objectives trained on different imbalanced datasets with different masking methods. Highest score and scores no less than one percent of the highest across each class imbalanced ratio are **bolded**.

Control: Depression	Standard BERT fine-tune	BERT fine-tune with reconstruction objective							AVG ± SD
		Random	Lexicon	Log-odds	TF-IDF	Sum Att	Top Att	NN	
1	**78.05**	**77.74**	**77.94**	**77.87**	**77.87**	**78.39**	**78.00**	**78.00**	**77.98 ± 0.20**
2	**71.10**	69.81	69.70	69.81	69.70	69.81	69.70	69.81	69.93 ± 0.06
4	**60.03**	**59.74**	**59.65**	**59.53**	**59.62**	**59.53**	**59.62**	**59.74**	59.68 ± 0.09
6	**54.83**	50.06	50.13	50.13	50.00	50.19	50.13	50.00	50.68 ± 0.07
8	**55.00**	52.07	48.74	48.38	48.44	52.21	48.07	48.56	50.18 ± 1.82
10	**52.20**	47.52	46.13	45.20	48.99	44.95	45.14	45.54	46.96 ± 1.51
AVG ± SD	**61.87 ± 10.40**	59.49 ± 12.06	59.49 ± 12.06	58.49 ± 13.07	59.10 ± 12.36	59.18 ± 12.76	58.44 ± 13.15	58.61 ± 13.03	

Table 4
F1-scores of BERT fine-tune with MKR loss trained on different imbalanced datasets with different masking methods. Highest score and scores no less than one percent of the highest across each class imbalanced ratio are **bolded**.

Control: Depression	Standard BERT fine-tune	BERT fine-tune with MKR only							AVG ± SD
		Random	Dep lexicon	Log-odds	TF-IDF	Sum Att	Top Att	NN	
1	**78.05**	**78.07**	**78.07**	**78.07**	**78.07**	**78.07**	**78.07**	**78.07**	**78.07± 0.00**
2	**71.10**	69.70	69.70	**71.32**	69.70	**71.43**	**71.43**	69.59	70.41 ± 0.93
4	**60.03**	**60.00**	**60.38**	**60.18**	**60.06**	**60.27**	**60.15**	**60.15**	60.17± 0.14
6	**54.83**	50.31	50.06	50.25	50.25	50.31	50.12	49.94	50.18± 0.10
8	**55.00**	50.37	50.25	50.31	50.64	50.31	50.64	50.64	50.45± 0.18
10	**52.20**	46.00	46.52	46.90	46.77	46.88	46.77	46.90	46.68 ± 0.34
AVG ± SD	**61.87 ± 10.40**	59.08 ± 12.63	59.16 ± 12.59	59.50 ± 12.76	59.25 ± 12.45	59.55 ± 12.45	59.53 ± 12.77	59.21 ± 12.77	

Table 5
F1-scores of BERT fine-tune with MER loss trained on different imbalanced datasets with different masking methods. Highest score and scores no less than one percent of the highest across each class imbalanced ratio are **bolded**.

Control: Depression	Standard BERT fine-tune	BERT fine-tune with MER only							AVG ± SD
		Random	Dep lexicon	Log-odds	TF-IDF	Sum Att	Top Att	NN	
1	**78.05**	**77.87**	**77.74**	**77.87**	**77.87**	**77.87**	**77.87**	**77.67**	**77.82± 0.05**
2	**71.10**	69.80	69.80	**69.80**	69.80	**69.80**	**69.80**	69.80	69.80± 0.00
4	**60.03**	**60.06**	**60.34**	**60.71**	**60.62**	**60.06**	**60.15**	**57.26**	59.89± 0.28
6	**54.83**	49.94	50.06	50.06	49.63	49.88	49.88	50.06	49.93± 0.16
8	**55.00**	52.37	52.29	53.62	52.66	52.22	52.44	52.44	52.58± 0.52
10	**52.20**	46.59	47.48	47.62	48.99	47.40	45.87	49.19	47.59± 1.05
AVG ± SD	**61.87 ± 10.40**	59.44 ± 12.27	59.62 ± 12.04	59.95 ± 11.90	59.93 ± 11.82	59.54 ± 11.71	59.33 ± 12.13	59.40 ± 12.42	

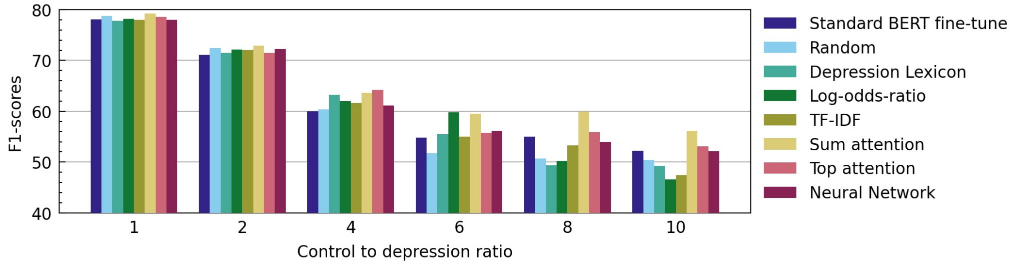

Figure 3
F1-scores of BERT further pre-train + fine-tune across masking methods and class imbalanced ratios.

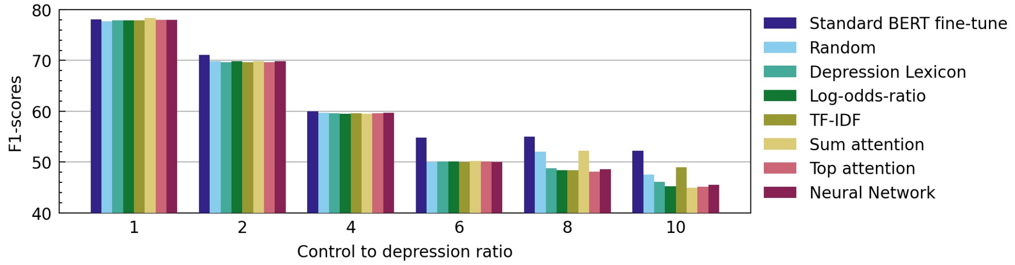

Figure 4
F1-scores of BERT fine-tune with reconstruction objective across masking methods and class imbalanced ratios.

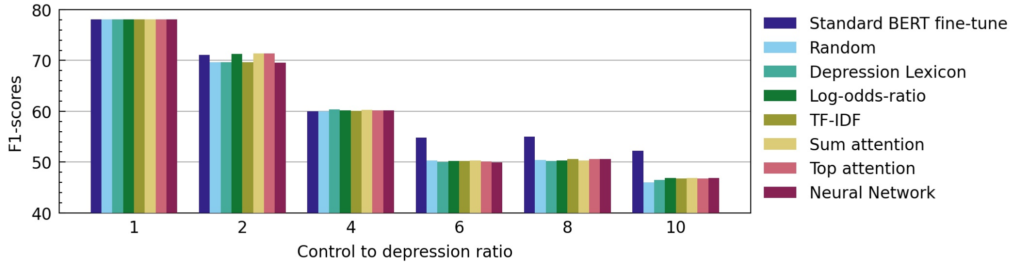

Figure 5
F1-scores of BERT fine-tune with MKR across masking methods and class imbalanced ratios.

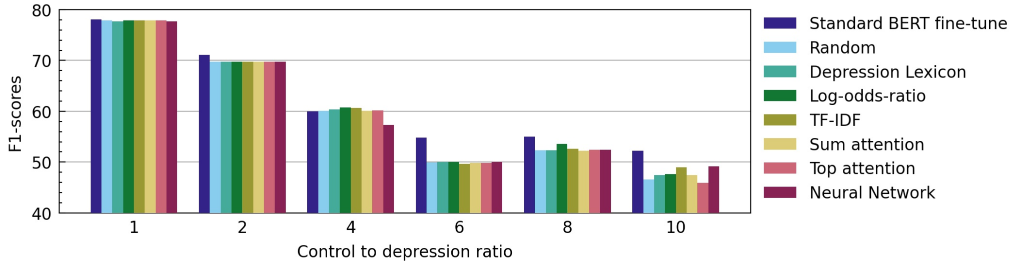

Figure 6
F1-scores of BERT fine-tune with MER across masking methods and class imbalanced ratios.

Table 6
Part of speech of words in each masked word list.

POS	Lexicon	Log-odds	TF-IDF	NN	Top att	Sum att
ADJ	9.82%	6.40%	8.63%	7.15%	6.73%	3.33%
ADP	0.00%	0.23%	0.30%	0.26%	0.23%	0.27%
ADV	0.00%	2.40%	3.03%	2.49%	2.29%	1.60%
CCONJ	0.00%	0.03%	0.00%	0.04%	0.04%	0.13%
DET	0.00%	0.03%	0.07%	0.10%	0.08%	0.17%
NOUN	43.75%	22.67%	31.37%	24.97%	23.88%	25.70%
PROPN	24.11%	57.00%	41.17%	28.69%	32.36%	35.40%
VERB	13.39%	6.83%	12.67%	13.41%	11.99%	2.70%
WordPiece	–	–	–	19.20%	17.99%	26.97%
Others	8.93%	4.40%	2.77%	3.69%	4.41%	3.73%

random masking, which performed over 4% worse than the best performer. In the most challenging class imbalanced ratio of 10, the best performer was also sum attention. As expected, the scores declined as the ratio of class disparity increased.

In contrast, when examining the BERT fine-tune with reconstruction objective training approach, BERT fine-tune with MKR, and BERT fine-tune with MER (see Tables 3–5), there was no discernible pattern, meaning that there was no obvious winner across masking methods, with one way sometimes outperforming the other. In fact, the basic BERT fine-tune model outperformed all other models trained using these three strategies.

4.2 Training Methods

Comparing the training approaches to standard BERT revealed that the BERT further pre-train method outperformed the others, followed by standard BERT, while BERT fine-tune with reconstruction objective approach, BERT fine-tune with MKR, and BERT fine-tune with MER performed worst with no discernible difference across these three approaches (see Tables 2–5). This trend generally held true for all class imbalanced ratios.

4.3 Class Imbalanced Ratios

Considering the BERT further pre-train + fine-tune training approach (see Table 2 and Figure 3), as expected, the scores declined as the ratio of class disparity increased. In the most difficult class imbalanced ratio of 10, the best result was achieved with sum attention.

Regarding BERT fine-tune with reconstruction objective training approach, BERT fine-tune with MKR, and BERT fine-tune with MER (see Tables 3–5), aside from the fact that the scores generally decreased as the ratio grows, there was no other noticeable pattern.

4.4 Masked Words

The full lists of the selected masked words can be found in our GitHub. To improve interpretability, we explored the masked words by identifying the parts of speech (POS) using SpaCy (Montani et al. 2020) and Linguistic Inquiry and Word Count[3] (LIWC)

[3] from LIWC2007 dictionary.

categories of the masked words from each method then calculating the percentage of the words in each category (see Table 6 and Table 7).

We have chosen to use these two features to study our masked words because POS was considered in the making of the depression lexicon (Losada and Gamallo 2020) and

Table 7
LIWC categories of words in each masked word list.

Category	**Lexicon**	**Log-odds**	**TF-IDF**	**NN**	**Top att**	**Sum att**
Linguistic Processes						
Total function words	3.57%	1.37%	1.27%	1.73%	1.55%	1.23%
Total Pronoun	3.57%	0.50%	0.23%	0.28%	0.25%	0.37%
Personal Pronouns	2.68%	0.17%	0.20%	0.15%	0.13%	0.23%
1st pers singular	0.00%	0.07%	0.00%	0.03%	0.03%	0.10%
1st per plural	0.00%	0.00%	0.03%	0.03%	0.02%	0.00%
2nd person	0.00%	0.03%	0.10%	0.03%	0.03%	0.10%
3rd pers singular	2.68%	0.07%	0.07%	0.04%	0.03%	0.03%
3rd pers plural	0.00%	0.00%	0.00%	0.02%	0.02%	0.00%
Impersonal pronouns	0.89%	0.33%	0.03%	0.13%	0.11%	0.13%
Common verbs	4.46%	0.70%	0.90%	1.14%	0.98%	0.27%
Past tense	0.89%	0.30%	0.43%	0.56%	0.48%	0.07%
Present tense	3.57%	0.40%	0.40%	0.51%	0.44%	0.10%
Psychosocial Processes						
Social processes	16.96%	2.70%	3.60%	2.68%	2.35%	2.83%
Family	0.89%	0.53%	0.43%	0.26%	0.24%	0.53%
Friends	0.89%	0.27%	0.30%	0.21%	0.20%	0.40%
Humans	3.57%	0.50%	0.20%	0.27%	0.26%	0.50%
Affective processes	25.00%	4.03%	8.27%	6.24%	5.48%	3.27%
Positive emotion	15.18%	1.67%	3.53%	2.96%	2.65%	1.87%
Negative emotion	9.82%	2.33%	4.70%	3.24%	2.79%	1.40%
Anxiety	3.57%	0.47%	1.03%	0.68%	0.57%	0.23%
Anger	2.68%	1.00%	2.03%	1.30%	1.11%	0.67%
Sadness	1.79%	0.23%	0.70%	0.64%	0.56%	0.23%
Cognitive Processes	8.04%	2.30%	5.53%	5.19%	4.38%	1.43%
Perceptual processes	3.57%	1.33%	2.17%	1.93%	1.75%	1.37%
See	0.89%	0.57%	0.67%	0.62%	0.58%	0.57%
Hear	1.79%	0.23%	0.50%	0.40%	0.39%	0.30%
Feel	0.89%	0.37%	0.53%	0.58%	0.50%	0.33%
Biological processes	25.89%	4.30%	5.37%	2.55%	2.36%	2.03%
Body	0.89%	1.13%	1.53%	0.89%	0.82%	0.83%
Health	21.43%	2.23%	2.57%	0.89%	0.84%	0.57%
Sexual	1.79%	0.67%	0.80%	0.34%	0.31%	0.77%
Ingestion	1.79%	0.80%	0.90%	0.59%	0.54%	0.27%
Relativity	2.68%	2.33%	4.53%	4.67%	4.26%	3.87%
Motion	0.89%	0.63%	1.30%	1.32%	1.17%	0.93%
Space	0.00%	0.67%	1.13%	1.61%	1.56%	1.37%
Time	1.79%	1.00%	2.07%	1.66%	1.47%	1.50%
Personal Concerns						
Work	0.89%	1.53%	3.50%	2.82%	2.62%	3.27%
Achievement	1.79%	0.57%	2.03%	1.95%	1.71%	1.17%
Leisure	5.36%	1.33%	2.07%	1.56%	1.50%	3.90%
Home	1.79%	0.37%	0.93%	0.55%	0.52%	0.40%
Money	0.00%	0.63%	1.73%	1.19%	1.04%	1.33%
Religion	8.04%	0.57%	1.07%	0.85%	0.90%	1.80%
Death	0.89%	0.23%	0.43%	0.33%	0.32%	0.33%
Others	11.61%	3.83%	7.83%	7.55%	6.43%	3.33%

LIWC has been commonly analyzed in studies related to language use of depression and depression detection in social media (Coppersmith, Dredze, and Harman 2014; Rude, Gortner, and Pennebaker 2004; Stirman and Pennebaker 2001; Yates, Cohan, and Goharian 2017; Cohan et al. 2018; Loveys et al. 2018; Nalabandian and Ireland 2019; Eichstaedt et al. 2018).

Regarding the top attention and neural network methods, they did not provide a list of masked words; thus, we applied masking models to the training set of our in-domain dataset and collected all masked terms indicated by the models.

Proper nouns (PROPN), nouns (NOUN), verbs (VERB), and adjectives (ADJ) were the top POS with the highest percentages across all masked methods. Proper nouns were the majority of the TF-IDF and log-odds-ratio's masking words.

Sum attention had the lowest proportion of verbs, adjectives, and adverbs when compared with other techniques (excluding depression lexicon). This similar pattern can be seen in LIWC, where sum attention had the fewest terms in the *common verbs*, *past tense*, and *present tense* categories.

In contrast, sum attention has the highest proportion of coordinating conjunctions and determinants when compared with other techniques.

For LIWC, *affective processes*, *social processes*, *negative emotion*, *positive emotion*, *cognitive processes*, *biological processes*, and *relativity* were among the most prominent categories across all masking methods.

Sum attention had the lowest proportion of words in the *affective processes*, *negative emotion*, *cognitive processes*, *biological processes*, *health*, *ingestion*, *anxiety*, *anger*, and *sadness* categories when compared with other methods. In contrast, sum attention had a greater proportion of *work*, *leisure*, and *1st person singular pronouns* than other methods.

5. Discussion

We discuss the masking methods, masked words, training methods, class imbalanced ratios, comparison to previous work, and, lastly, limitations and future work.

5.1 Masking Methods

In the BERT further pre-train + fine-tune training approach, selective masking methods outperform random masking methods (see Table 2). This confirms some of our initial thoughts that selecting specific words to mask would be better at guiding the model to learn the language patterns related to depression than randomly masking. Among selective masking methods, sum attention outperformed other techniques, followed closely by top attention and neural network. Here we further discuss the possible reasons behind it.

First, while the other methods only select whole words as masked words, the only three masked words selection methods that mask WordPiece tokens (e.g., '##any', '##uld', '##and', '##ing') are sum attention, neural network, and top attention, which are also the three best performing methods. Because sum attention has the highest percentage of WordPiece tokens in the masked word list (see Table 6) and is also the best performer, it is possible that masking WordPiece tokens might help the model better learn domain-specific words outside the vocabulary of BERTtokenizer, such as abbreviations or casual text commonly used in social media.

We originally anticipated that top attention and neural networks would outperform other approaches that use static masked word sets. This is because we expected that these approaches would be able to learn the context-specific linguistic patterns for each sequence. Surprisingly, though, sum attention outperformed these approaches. A potential explanation is that gathering data from all samples in the dataset, as in sum attention, enables the model to acquire more consistent patterns for word masking. Sum attention is analogous to using corpus statistics, while top attention and neural networks utilize sample statistics, which may not catch the consistent signal in its whole.

For example, in neural network, we observed that in some samples the masked words were aggregated in the latter part of the sequence. This is possibly due to its important words selection algorithm, that is, if many words in the first part of the sequence are not marked as important, we would have many words left in the buffer. This would then cause the buffer to have very similar words to the original sequence. Consequently, when words from the last part of the sequence are added to the buffer, there is a possibility that the probability likelihood would be very close to the original sequence. Hence, many of the words in the last part of the sequence could be marked as important.

Log-odds-ratio and TF-IDF are corpus-based frequency-based masked words selection techniques. In comparison to sum attention, top attention, and neural network, they fared substantially poorer. This is to be anticipated, given that common terms seldom correspond directly with important words. For example, when we first apply standard TF-IDF to our depression and control document, the most important words from each document were mostly stopwords and common words, which are very high in frequency. Obviously, we have attempted to remove stopwords and get just the terms unique to each document, but our efforts have also not yielded satisfactory results.

The depression lexicon was the only method that was based on a dictionary, and it was the second worst performer overall when trained with the BERT further pre-train + fine-tune approach. Several reasons can be deduced. First, this is likely because the lexicon has only 112 words, much smaller when compared with other methods that have more than 3,000 words. Second, the topics of these words are also quite limited, as seen in Table 7, that is, more than half of the words from the lexicon are only related to *affective processes* and *biological processes*. Last, the lexicon was created from text sources that do not include Reddit, which means the lexicon might not be tailored to the language used in this RSDD dataset.

5.2 Training Methods

BERT further pre-training performed the best, followed by standard BERT fine-tune and BERT fine-tune with reconstruction objective approach, respectively (see Tables 2 and 3).

The superior performance of further pre-training is intuitive, since helping the model to initially learn the broad language patterns on the real downstream job aids the model's performance at the fine-tune stage.

In any case, it is strange to us that BERT fine-tune with additional reconstruction objectives as well as BERT fine-tune with MKR and BERT fine-tune with MER did not outperform the standard BERT fine-tune approach (see Tables 3–5). This might be due to the difference between our method and the original work (Moon et al. 2021), which are the number of selected masked words (ours = 3,000 and theirs = 20) and the independent probability that each word is masked (ours $p = 0.9$, theirs $p = 0.5$). These differences might cause the number of masked words in each sequence for the MKR technique in our work (average between 9 and 104 tokens) to be much higher than the original work. It is possible that when there is more than one objective to train at once in the fine-tuning phase, masking too many tokens in the sequences might affect the ability of the model. In any case, further studies exploring all the possible configurations may be required.

5.3 Class Imbalanced Ratios

As anticipated, the F1-scores declined for all three training approaches as the datasets became more unbalanced owing to the rising complexity of the task (see Figures 3 to 6). This helps reemphasize the difficulty of modeling in domains such as depression, where class imbalance is common.

For both the BERT further pre-train + fine-tune and the BERT fine-tune with reconstruction objective training approaches, the performance of various masking methods at smaller imbalanced ratios is almost identical. A likely explanation is that the signal of depression is readily discernible at low imbalanced ratios. Even though the models were trained using a variety of masking techniques, they were all able to identify depression with comparable accuracy. This also demonstrated why, in the domain of depression, it is crucial to include imbalanced ratios into research and not be deceived by strong performance at a 1:1 or small imbalanced ratio.

Looking at the further pre-training approach (see Figure 3) where our masking method worked best, sum attention performed best, even at a larger ratio, than other methods. This suggests the robustness of sum attention. Top attention, standard BERT, and neural network followed closely after. On the other hand, when the ratio is large, all frequency-based approaches, including log-odds-ratio, TF-IDF, and depression lexicon, perform poorly. This outcome seems sensible, given that frequency-based techniques rapidly lose significance when the ratio is large, since frequency-based metrics become less effective when there is an imbalance across classes. Other methods, such as sum attention, may dynamically adapt more effectively (but still not amazing, since the accuracy is only around 55%).

5.4 Masked Words

First, it can be observed that the important words obtained from all masked words selection methods contain words from various LIWC categories and parts of speech (see Tables 6 and 7), some of which are also keywords directly or closely related to depression (e.g., “dosage,” “therapy,” “psychiatrist,” “mentally”). It is possible that by selectively masking these diverse important words altogether, the models are forced to learn the language patterns from their surrounding context, thus, reducing the model’s tendency to overly rely on depression keywords. This might be one of the reasons for the enhanced performance when further pre-training the model on selectively masked dataset.

Next, we analyze the masked words of sum attention according to the POS tags and LIWC.

5.4.1 POS Tags. Based on the findings (see Table 6), the proportion of adjectives and adverbs in the sum attention masked word list is lower than other methods. It’s likely that including these parts of speech in the context text might assist the model to better comprehend the context required to predict the masked words. Previous research has shown that users with depression use more adverbs in their Reddit posts (Bucur, Podină, and Dinu 2021) and that adjectives are a useful component of the depression lexicon for identifying depression (Losada and Gamallo 2020).

Moreover, sum attention has the highest percentage of coordinating conjunctions (CCONJ). It is thus possible that masking these stopwords enhance the performance of the model. For example, it is possible that by masking the coordinating conjunction out (e.g., "and," "or," "but") the language model would be forced to capture the semantics of the surrounding words/phrases/clauses to be able to predict the correct conjunction that links them together. This resonates with previous work showing that language models can understand the conjunction of facts expressed by the word "and" (Talmor et al. 2020).

The log-odds-ratio and TF-IDF are the two methods that contain the highest percentage of proper nouns in their masked words as well as having proper nouns as the majority of their masked words. We observed that their masked words indeed include many specific names or words such as "tonberrys," "jizzlam," "kamenev," "miggy," "daniel," and "bourgeoisie." It is possible that, in our task, masking these proper nouns may not guide the model to learn the context beneficial for depression detection, hence the lower performance seen with log-odds-ratio and TF-IDF. This might be due to the observation that individuals with depression tend to use fewer proper nouns in comparison with control users (Bucur, Podină, and Dinu 2021), which is due to their lower interest in people and objects (De Choudhury et al. 2016).

5.4.2 LIWC. Next, we explore the LIWC counts to analyze the components of the masked words (see Table 7). From the LIWC counts, we can see that for all methods, the majority of masked words are from the categories related to *affective*, *biological*, *social*, and *cognitive processes*. However, for sum attention, the words in these four categories are not as high in percentage as other methods. In addition, it is noticeable that sum attention contains higher percentages of masked words in the *personal concerns* category, namely, *work* and *leisure*. It is possible that other than *affective processes*, *biological processes*, *social processes*, and *cognitive processes*, the depressed users also mentioned a lot about *work* and *leisure*. This is probably the reason why when words from all of these categories are masked in the sum attention method the model can better capture the depression language and yield better results than other masking methods, which do not include as many words from the *work* and *leisure* categories. These results resonate with previous work (Zhang et al. 2021; De Choudhury et al. 2013; Coppersmith, Dredze, and Harman 2014), which have proven that the depression group significantly used more words in the *anxiety*, *anger*, *swear*, and *negative emotion* categories than the non-depression group. Moreover, De Choudhury et al. (2013) reported that words related to treatment, relationships, life, and disclosure were found in the depressed group at a high frequency. When observed, the words in their results are indeed very similar to the *biological processes*, *life*, and *leisure* categories in our results. Bucur and Dinu (2020) also confirmed that depressed users are generally more focused on topics about their personal life experience and sentiments.

Another observation is that sum attention has the highest percentage of *1st person singular pronouns* ("I", "me," "mine") in its masked word list while having less or equal percentage of other pronouns (*1st person plural, 2nd person , 3rd person singular, 3rd person plural, and impersonal pronouns*) than other methods. In our case, by masking more first person singular pronouns, the model might be able to better learn the context related to the users themselves, which might also contain information about their mental states. This might also be related to previous works (Coppersmith, Dredze, and Harman 2014; Stirman and Pennebaker 2001; Cohan et al. 2018; Morales, Scherer, and Levitan 2018) that have proven that depressed users use significantly more words in the *1st person singular pronouns* category, which indicates higher self-preoccupation, and significantly fewer words in *1st person plural*, which suggests self-focused attention in depressed individuals (De Choudhury et al. 2016).

5.5 Comparison to Previous Work

Because of the changes made to the RSDD datasets by this study, we should be extremely cautious when comparing our results to other works.

The first thing that makes our work different from what has been done before is that we are the first to try to make a depression detection dataset that is both filtered by content (words about mental health had been taken out of the depression group of the RSDD dataset) and by time (using the Time-RSDD dataset, we only took users who have their condition right now). The ratio of control-to-depression groups in the dataset is also another factor that needs to be kept in mind when comparing with past work. Here we compare the performance of our models with similar previous work.

Our work would be the most similar to Zhang et al. (2021), who also used a transformer-based classification model to detect depression using chunks of text. Note that their dataset contained an equal number of depressed and control users. Also note that their work was done on Twitter, while ours was done on Reddit. Their RoBERTa model achieved an F1-score of 78.0% while our BERT further pre-train + fine-tune model with sum attention masking achieved a slightly higher F1-score of 79.25% at the same ratio.

For the work done on the RSDD dataset, Yates, Cohan, and Goharian (2017) achieved an F1-score of 51% with a user-CNN model, Rao et al. (2020) achieved an F1-score of 54% with their multi-gated LeakyReLU CNN, and Song et al. (2018) used a feature-based attention model and was able to achieve an F1-score of 56%. Note that these results were obtained at the user-level from the original RSDD dataset, which has a control-to-depression user ratio of 1:12. As for our dataset, the results were obtained at the chunk-level at a ratio of 1:10, which achieved an F1-score of 56.20% with sum attention.

5.6 Limitations and Future Work

The limitations of this study are that this study only focused on the RSDD dataset. However, this dataset has many advantages. First, high-precision matching patterns and human annotation were used to eliminate false-positive self-declared posts. Second, it is one of the few datasets that attempted to eliminate words related to mental health. Third, temporal information about the onset of depression for some users is available in the Time-RSDD dataset. Lastly, it is one of the largest datasets on depression detection.

We extracted the masked words from a balanced dataset because we do not want class bias in the masked word lists. Then, this same set of masked words were used to selectively mask larger datasets with higher imbalanced ratios. The results demonstrated that sum attention consistently performed better than other selective masking methods even at higher class imbalanced ratios which implies that we might be able to apply the same set of masked words on more imbalanced datasets with minimal performance degradation. However, we believe that extraction of important words from a larger, more imbalanced dataset or dataset from a different domain as well as the application of the obtained masked words to other datasets could also be explored.

In this study, we focused on comparing the existing training and masking methods proposed in previous work. However, we suggest that other training methods are possible. For example, because the results have shown that attention-based masking methods and methods involving masking models performed relatively well, we can also try creating an end-to-end model that can learn to select the masked words at the same time as training the model for depression detection.

We investigated masked words with POS and LIWC. However, it would be very interesting to study other language patterns, which can help link the principles learned in the healthcare domain with the field of machine learning and deep learning.

6. Conclusion

The original premise of this study is that depression identification in text is a complex phenomenon. It might appear that the model performed well because it was just trained on a depressed text dataset. The model, however, frequently manipulates performance by excessively relying on depression keywords, so it was unable to accurately represent the inherent linguistic nature of depressive posts. As a result, a series of comparisons were made in this study to assess how well different selective masking techniques compared to random masking. We showed that selective masking performs better than random masking in most situations. We determined summation of attention to be the best performing and most robust selective masking technique. Our findings also showed that reconstructing the masked words is more favorable during the pre-training phase than it is during the fine-tuning phase. Our research validated selective masking as a promising method for reducing keyword bias and injecting knowledge of the downstream task into the model.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

AlSagri, Hatoon S. and Mourad Ykhlef. 2020. Machine learning-based approach for depression detection in Twitter using content and activity features. *IEICE Transactions on Information and Systems*, 103(8):1825–1832. `https://doi.org/10.1587/transinf.2020EDP7023`

Bucur, Ana Maria and Liviu P. Dinu. 2020. Detecting early onset of depression from social media text using learned confidence scores. *arXiv preprint arXiv:2011.01695*. `https://doi.org/10.4000/books.aaccademia.8305`

Bucur, Ana Maria, Ioana R. Podină, and Liviu P. Dinu. 2021. A psychologically informed part-of-speech analysis of depression in social media. *arXiv preprint arXiv:2108.00279*. `https://doi.org/10.26615/978-954-452-072-4_024`

Chalkidis, Ilias, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*. `https://doi.org/10.18653/v1/2020.findings-emnlp.261`

Cohan, Arman, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. SMHD: A large-scale resource for exploring online language usage for multiple mental health conditions. *arXiv preprint arXiv:1806.05258*.

Coppersmith, Glen, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60. `https://doi.org/10.3115/v1/W14-3207`

Coppersmith, Glen, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39. `https://doi.org/10.3115/v1/W15-1204`

De Choudhury, Munmun, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Seventh International AAAI Conference on Weblogs and Social Media* , pages 128–137.

https://doi.org/10.1609/icwsm.v7i1.14432

De Choudhury, Munmun, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2098–2110. https://doi.org/10.1145/2858036.2858207, PubMed: 29082385

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eichstaedt, Johannes C., Robert J. Smith, Raina M. Merchant, Lyle H. Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A. Asch, and H. Andrew Schwartz. 2018. Facebook language predicts depression in medical records. In *Proceedings of the National Academy of Sciences*, 115(44):11203–11208. https://doi.org/10.1073/pnas.1802331115, PubMed: 30322910

Gu, Yu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23. https://doi.org/10.1145/3458754

Gu, Yuxian, Zhengyan Zhang, Xiaozhi Wang, Zhiyuan Liu, and Maosong Sun. 2020. Train no evil: Selective masking for task-guided pre-training. *arXiv preprint arXiv:2004.09733*. https://doi.org/10.18653/v1/2020.emnlp-main.566

Gururangan, Suchin, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*. https://doi.org/10.18653/v1/2020.acl-main.740

Jamil, Zunaira, Diana Inkpen, Prasadith Buddhitha, and Kenton White. 2017. Monitoring tweets for depression to detect at-risk users. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, pages 32–40. https://doi.org/10.18653/v1/W17-3104

Kawintiranon, Kornraphop and Lisa Singh. 2021. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4725–4735. https://doi.org/10.18653/v1/2021.naacl-main.376

Lin, Chen, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2021. EntityBERT: Entity-centric masking strategy for model pretraining for the clinical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 191–201. https://doi.org/10.18653/v1/2021.bionlp-1.21

Lin, Chenhao, Pengwei Hu, Hui Su, Shaochun Li, Jing Mei, Jie Zhou, and Henry Leung. 2020. SenseMood: Depression detection on social media. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 407–411. https://doi.org/10.1145/3372278.3391932

Losada, David E. and Fabio Crestani. 2016. A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–39. https://doi.org/10.1007/978-3-319-44564-9_3

Losada, David E., Fabio Crestani, and Javier Parapar. 2018. Overview of eRisk 2018: Early risk prediction on the internet (extended lab overview). In *Proceedings of the 9th International Conference of the CLEF Association, CLEF*, pages 1–20. https://doi.org/10.1007/978-3-319-98932-7_30

Losada, David E. and Pablo Gamallo. 2020. Evaluating and improving lexical resources for detecting signs of depression in text. *Language Resources and Evaluation*, 54(1):1–24. https://doi.org/10.1007/s10579-018-9423-1

Loveys, Kate, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith. 2018. Cross-cultural differences in language markers of depression online. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 78–87. https://doi.org/10.18653/v1/W18-0608

MacAvaney, Sean, Bart Desmet, Arman Cohan, Luca Soldaini, Andrew Yates, Ayah Zirikly, and Nazli Goharian. 2018. RSDD-time: Temporal annotation of self-reported mental health diagnoses. *arXiv preprint arXiv:1806.07916*. https://doi.org/10.18653/v1/W18-0618

Montani, Ines, Matthew Honnibal, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, Henning Peters, Maxim Samsonov, Jim Geovedi, Jim Regan, György Orosz, Paul O'Leary McCann, Søren Lind Kristiansen, Duygu Altinok, Roman, Leander Fiedler, Grégory Howard, Wannaphong Phatthiyaphaibun, Explosion Bot, Sam Bozek, Mark Amery, Yohei Tamura, Björn Böing, Pradeep Kumar Tippa, Leif Uwe Vogelsang, Ramanan Balakrishnan, Vadim Mazaev, Greg Dubbin, Jeanne Fukumaru, Jens Dahl Møllerhøj, and Avadh Patel. 2020. *explosion/spaCy: v2.3.5: Bug fixes and simpler source installs*. Zenodo. https://doi.org/10.5281/zenodo.4317367

Moon, Seung Jun, Sangwoo Mo, Kimin Lee, Jaeho Lee, and Jinwoo Shin. 2021. MASKER: Masked keyword regularization for reliable text classification. In *AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v35i15.17601

Morales, Michelle, Stefan Scherer, and Rivka Levitan. 2018. A linguistically-informed fusion approach for multimodal depression detection. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 13–24. https://doi.org/10.18653/v1/W18-0602

Nalabandian, Taleen and Molly Ireland. 2019. Depressed individuals use negative self-focused language when recalling recent interactions with close romantic partners but not family or friends. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 62–73. https://doi.org/10.18653/v1/W19-3008

Orabi, Ahmed Husseini, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of Twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97. https://doi.org/10.18653/v1/W18-0609

Rao, Guozheng, Yue Zhang, Li Zhang, Qing Cong, and Zhiyong Feng. 2020. MGL-CNN: A hierarchical posts representations model for identifying depressed individuals in online forums. *IEEE Access*, 8:32395–32403. https://doi.org/10.1109/ACCESS.2020.2973737

Rude, Stephanie, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133. https://doi.org/10.1080/02699930441000030

Sekulić, Ivan and Michael Strube. 2020. Adapting deep learning methods for mental health prediction on social media. *arXiv preprint arXiv:2003.07634*. https://doi.org/10.18653/v1/D19-5542

Shen, Guangyao, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, pages 3838–3844. https://doi.org/10.24963/ijcai.2017/536

Song, Hoyun, Jinseon You, Jin-Woo Chung, and Jong C. Park. 2018. Feature attention network: interpretable depression detection from social media. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*.

Stirman, Shannon Wiltsey and James W. Pennebaker. 2001. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine*, 63(4):517–522. https://doi.org/10.1097/00006842-200107000-00001, PubMed: 11485104

Sun, Chi, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. https://doi.org/10.1007/978-3-030-32381-3_16

Talmor, Alon, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758. https://doi.org/10.1162/tacl_a_00342

Tian, Hao, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. *arXiv preprint arXiv:2005.05635*. https://doi.org/10.18653/v1/2020.acl-main.374

Wołk, Agnieszka, Karol Chlasta, and Paweł Holas. 2021. Hybrid approach to detecting symptoms of depression in social media entries. *arXiv preprint arXiv:2106.10485*.

Wolohan, J. T., Misato Hiraga, Atreyee Mukherjee, Zeeshan Ali Sayyed, and Matthew Millard. 2018. Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP. In *Proceedings of the First International*

Workshop on Language Cognition and Computational Models, pages 11–21.

Yates, Andrew, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*. `https://doi.org/10.18653/v1/D17-1322`

Zhang, Yipeng, Hanjia Lyu, Yubao Liu, Xiyang Zhang, Yu Wang, Jiebo Luo, et al. 2021. Monitoring depression trends on Twitter during the COVID-19 pandemic: Observational study. *JMIR Infodemiology*, 1(1):e26769. `https://doi.org/10.2196/26769`, PubMed: 34458682

Zogan, Hamad, Imran Razzak, Shoaib Jameel, and Guandong Xu. 2021. DepressionNet: A novel summarization boosted deep framework for depression detection on social media. *arXiv preprint arXiv:2105.10878*.