Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5418–5427
Marseille, 20-25 June 2022

Leveraging Mental Health Forums for User-level Depression Detection on Social Media

Sravani Boinepelli[1], Tathagata Raha[1], Harika Abburi[1], Pulkit Parikh[1], Niyati Chhaya[2], Vasudeva Varma[1]
[1]IIIT-Hyderabad
Information Retrieval and Extraction Lab, Gachibowli, Hyderabad, Telangana, India
sravani.boinepelli,tathagata.raha,harika.a,pulkit.parikh@research.iiit.ac.in, vv@iiit.ac.in
[2]Adobe Research
Bengaluru, Karnataka, India
nchhaya@adobe.com

Abstract

The number of depression and suicide risk cases on social media platforms is ever-increasing, and the lack of depression detection mechanisms on these platforms is becoming increasingly apparent. A majority of work in this area has focused on leveraging linguistic features while dealing with small-scale datasets. However, one faces many obstacles when factoring into account the vastness and inherent imbalance of social media content. In this paper, we aim to optimize the performance of user-level depression classification to lessen the burden on computational resources. The resulting system executes in a quicker, more efficient manner, in turn making it suitable for deployment. To simulate a platform agnostic framework, we simultaneously replicate the size and composition of social media to identify victims of depression. We systematically design a solution that categorizes post embeddings, obtained by fine-tuning transformer models such as RoBERTa, and derives user-level representations using hierarchical attention networks. We also introduce a novel mental health dataset to enhance the performance of depression categorization. We leverage accounts of depression taken from this dataset to infuse domain-specific elements into our framework. Our proposed methods outperform numerous baselines across standard metrics for the task of depression detection in text.

Keywords: Depression Detection, Text classification, Semi-supervised, weakly-supervised and unsupervised learning, Social Media Processing

1. Introduction

Depression is viewed as the largest contributor to global disability and often co-occurs with anxiety or other psychological and physical disorders. One of the many risks of clinical depression is suicide. Research has indicated that approximately two-thirds of people who die by suicide were dealing with depression at the time of death (Richards and OHara, 2014). Meanwhile, according to the World Health Organization, more than 300 million people from varied demographics suffer from depression and nearly 50% of people worldwide with clinical depression remain untreated (WHO, 2017).

Over the last couple of years, there has been an influx of depression and mental health posts on social media which has only been exacerbated by the COVID-19 pandemic. The abundance of abusive language on the web has also contributed to the profusion of mental health and depression-related issues, especially in teens, making life online increasingly hazardous (Fersko, 2018; Karim et al., 2020). In this paper, we take steps towards identifying not just singular accounts of depression (post-level classification), but towards detection of individuals suffering from depression (user-level classification). By post-level classification, we refer to the labelling of isolated social media posts whereas user-level classification labels a user based on a collection of posts or post history of the user. Our goal is to be able to forecast and reach out to these users to provide them with help and resources before their condition escalates to suicidal ideation. However, this is a colossal task wrought with obstacles such as the inherent imbalance and vastness of social media.

Automated research efforts towards identifying and reaching out to social media users who suffer from depression are limited due to the scarcity of personal data a user is willing to disclose on public forums. Though awareness is being raised, there are still many countries in which social stigma and discrimination have prohibited people from getting proper treatment and support for clinical depression (Wainberg et al., 2017). As a result, victims often turn to less formal resources on the Internet, sharing their experiences and challenges through online forums, blogs, or subreddits that allow a certain level of anonymity. These platforms not only enable them to ask for advice and talk about their condition, but also allow them to develop a shared sense of community as they no longer have to face such problems alone. Social media has therefore become a valuable source of linguistic cues that could help identify depression from texts while retaining user privacy (Choudhury and De, 2014; Manikonda and Choudhury, 2017). These anonymous forums therefore tick all the boxes for our task.

Existing datasets in this space are often derived from social media platforms such as Twitter and Reddit. For our experiments, we also consider mental health forums as they have a similar structure, emulating the informal dialogue that is found on social media (e.g. mental health subreddits) while simultaneously using diction employed by formal mental health professionals. We therefore introduce a new dataset obtained from a collection of anonymous mental health posts. Incorporating such a domain-specific language resource into our architecture significantly increases the accuracy of the diagnosis made on these datasets.

We focus on Reddit as it is a unique platform on which users can choose to create "throwaway" accounts that are not associated with their main accounts. This allows them to feel safe to make posts or comments disclosing sensitive information without revealing their identity. We believe this simulates the way people suffering from depression typically share their experiences online. This also grants the added advantage of allowing longer post lengths in contrast to platforms such as Twitter and Instagram, thus being more helpful to our user-level classification task, which hugely relies on context. The Reddit Self-reported Depression Diagnosis (RSDD) dataset (Yates et al., 2017) was selected for this study as it simulated challenges one would face when dealing with the vastness of social media on limited resources, as well as the class imbalance of depressed vs other miscellaneous content.

Our key contributions can be summarized as follows:

- We design a hierarchical neural framework for identifying cases of depression at a user level.
- Our system efficiently handles issues raised when utilizing a large and class-imbalanced dataset such as RSDD to its full extent. This scope can be expanded to build models that would be deployed on social media platforms.
- We introduce a novel dataset derived from the mental health blogs to incorporate domain-specific knowledge into our framework.
- Our proposed methods yield better results than previous baselines, including many deep learning and machine learning methods.

2. Related Work

Early research has been predominantly theoretical or used fundamental statistical analysis on a limited amount of data sourced from interviews, and surveys (Homan et al., 2014; Li et al., 2016; Burgess, 2019). Li et al. (2016) is one such case study that formed its observations on the influence of Chinese society on mental health based on interviews conducted from the Chinese online depression community, SunForum. Choudhury et al. (2014) uses a series of statistical methods and measures activity, emotion, and linguistic style to predict the onset of post-partum depression in Facebook data. Their results are also corroborated by interviewing mothers suffering post-partum depression. However, it is difficult to expand such models to run on large amounts of data efficiently.

Though the scope has grown to include more enhanced NLP models and social media derived datasets, the amount of data used is often less, or restricted (Coppersmith et al., 2015; Losada and Crestani, 2016; Yates et al., 2017). Shared tasks in the field of mental health detection contribute significantly toward developing solutions and raising awareness for mental health. However, they frequently use small-scale and balanced datasets to conduct and evaluate experiments conveniently. Coppersmith et al. (2015) is such a shared task that sampled Twitter users belonging to three groups: depression, PTSD, and control. Participants were asked to binarily classify a user as depression or control, PTSD or control, and depression or PTSD. The dataset for each experiment was split for training and testing purposes and further balanced to reflect user population inaccurately. Nevertheless, this allowed them to circumvent limitations on data distribution from Twitter's terms of service. It also made classifier creation and interpretation easier.

Zogan et al. (2021) also utilizes Twitter data and proposes a novel deep learning framework for automatic depression detection consisting of a Convolutional Neural Network (CNN) coupled with attention-enhanced Gated Recurrent Units (GRU) models. While most work in this space identifies individual posts related to depression, this work focuses on detecting depressed social media users using their post history and selects relevant content using summarization strategies. Yates et al. (2017) is another piece of work that contributed to the detection of depressed users using text and language alone, similar to our problem statement. They also introduce a large-scale depression dataset composed of Reddit users with self-reported depression diagnoses matched with control users. However, they cut down a sizable amount of their data while running their models using arbitrary sampling methods. Through our experiments, we aim to deliver similar results while using much larger datasets that would simulate the amount of data and imbalance on social media.

Several studies analyze emotion, language style, visual data, user profile information and user activity online to use as features to identify depression (Shen et al., 2017; Shen et al., 2018; Asad et al., 2019; Zogan et al., 2021). This is done with the help of tools such as the LIWC(Pennebaker et al., 2015), which maps linguistic characteristics to categories and provides insights into a user's personality and health. Some notable linguistic features that are known to manifest in ”depressed language” are negatively-valenced words (Beck, 1983) and frequent use of first-person pronouns (Pyszczynski et al., 1987). These features are also taken into consideration while designing our cost-effective user-level depression detection system.

3. Datasets

3.1. Reddit Self-reported Depression Diagnosis (RSDD) dataset

We use the large-scale publicly available RSDD dataset (Yates et al., 2017), which contains posts of Reddit users who claim (and have been manually validated) to be diagnosed with depression. Users who were randomly selected and whose posts did not contain any depression-related keywords were categorized into the non-depressed group. The RSDD dataset contains 9,210 depressed users and 107,274 non-depressed users, with an average of 969 posts for each user and a median of 646. Labels were provided at the user level but not at the individual post level. So posts by a depressed user do not necessarily have to talk about depression alone.

Previous work that has been done on this dataset has largely cut down on the size by sampling sections of the dataset to increase efficiency. Our objective is to leverage all the data given to us to appropriately run experiments that replicate the size and inherent imbalance of depression to non-depressed data available online. However, there are certain obstacles we must overcome, especially if resources available to us are limited, in order to optimize time and increase efficiency without forfeiting user accuracy.

3.2. Mental Health Blog (MHB) Dataset

We also introduce a novel dataset[1], which includes posts written in English from public mental health forums, in order to add some domain-specific elements into our architecture. We explore such online forums because they perfectly tread the line between using less formal diction as seen in social media, while also discussing genuine accounts from those who think they may be suffering from, or have been diagnosed with, depression. To the best of our knowledge, this is the first work to introduce this dataset and leverage posts from these mental health forums. To collect data, we crawled mental health blogs for posts spanning a timeline from 31st December 2011 to 25th June 2020. This yielded a mental health language resource containing a total of 39248 posts with 9354 total unique users. The dataset also comes with other helpful information such as date of posting, number of posts previously posted by the author, number of likes received to mark how helpful a post was, etc. In this work, we explore only the posts made by each user.

Table 1 provides a series of noteworthy statistics for both RSDD and Mental Health Blog datasets to give us a better understanding of the contents of the dataset. Mental health blogs have more context in each post which helps to incorporate more domain-specific and linguistic cues to the model. Figure 1 shows the word cloud generated from this dataset. Here, the size of the font for each tokenized word corresponds to the frequency of use for that word. Since the words are tokenized, similar words such as ’feels’ and ’feeling’ are grouped together until only morphs of the word (or words that can stand alone such as ’feel’) remain. We also show a few posts from the Mental Health Blog dataset in Table 2.

Dataset	Statistic	Count
MHB	Mean length of posts	230
	# of Sentences	452756
	NER count	188259
	Total word count	4439028
	# of Unique words	49904
	Forms of Depression$^{\alpha}$	7708
	Depression WordNet$^{\beta}$	359
RSDD	Avg. # of posts/user	969
	Mean length of posts	148
	# of Depressed Users	9210
	# of Non-Depressed Users	107,274
	Vocabulary Size	966881

$^{\alpha}$ Includes count of inflected forms of the word depression such as depressing/depressed (which result in the same morpheme after lemmatization)

$^{\beta}$ Includes count of top words (selected from the WordNet for Depression) occuring in the dataset. eg. misery, sorrow

Table 1: Relevant Dataset Statistics

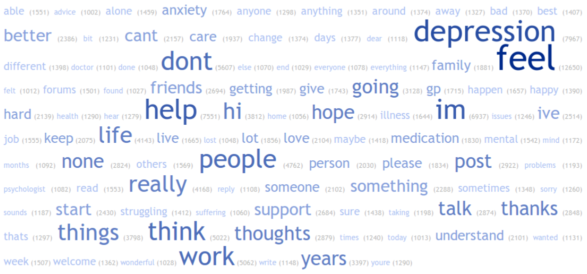

Figure 1: Word Cloud of posts from Mental Health Blog dataset

3.3. Ethical concerns

Social media data is often sensitive, and this is especially true when the data is related to mental health. Our Mental Health Blog dataset contains only publicly available posts. We are committed to following ethical practices, which include protecting the privacy and anonymity of the users. The author’s usernames, which could contain sensitive information related to the names or locations of the user, are not saved or used at all. Instead, the information was pre-processed and replaced with user IDs.

[1] Our dataset is publicly available: `https://bit.ly/3IgRMgD`

Post
I have been having bad thoughts about regret for the last month and I need some help im so stuck and feel like I want to die.
I hate my dogs I hate being a parent I hate being a widow I hate mess I hate my brain I hate my feelings I hate my false glimmer of hope I hate being me.
I don't do mood journals, don't know why, just something that doesn't click with me. I'm a big supporter of getting a good nights sleep, I believe it plays a huge role in mental health. They say you need an average of 8 hours of sleep a night, but I can only get 6 at most. Of that 6 hours I used to wake up 5/6 times a night. I've implemented a bit of a 'before bed' schedule to relax me before bed. I try to eat dinner at around 7 and I try to keep the portion size small. If I over eat especially meaty/fatty things I tend to get heart burn and generate a lot of body heat and i sleep terribly. At 9:30 my phone goes into night screen mode and I try to look at at as little as possible. I put on a light show or a movie, make a cup of peppermint tea and just zone out. 90% of the time i m relaxed enough to fall asleep within 10/20 minutes. I generally only wake up once a night. If i get a terrible night sleep, I forgive myself and tell myself there is always tonight. i have also done a bit of research into sleep hygiene to help things along.

Table 2: Example posts from Mental Health Blog dataset

4. Proposed Methodology

Our depression detection architecture can be divided into subsystems (as shown in Figure 2) that individually tackle various challenges and optimize time and resources in conjunction. Three major challenges one faces while designing a low-resource detection system that encompasses large-scale, social media-derived datasets include:

- memory and time constraints (to build an application that works in real-time on social media),
- customizing the model to single out depression content and,
- handling the class imbalance that comes with the vast diversity of topics on the Internet.

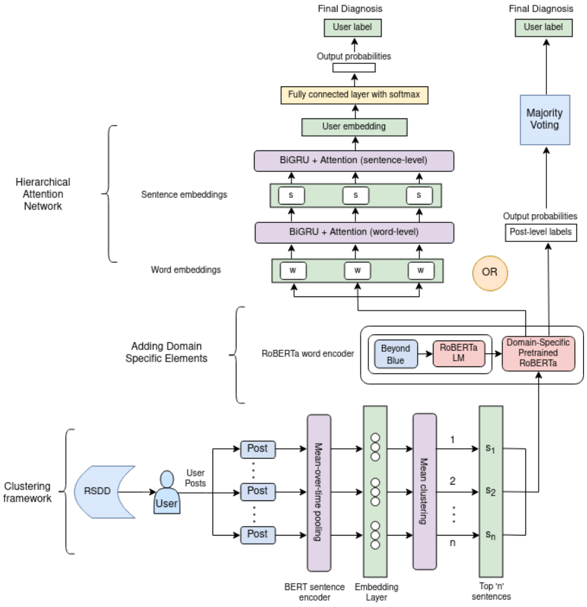

Figure 2: System Architecture

4.1. Clustering framework

The first issue addressed is leveraging the entire RSDD dataset while simultaneously dealing with time constraints and cutting down the cost of memory and compute resources. Work done on this dataset has often put limits on model inputs such as the number of posts and post length. Summarization could be used to reduce the load on memory (Zogan et al., 2021). However, this may be at the expense of losing key linguistic features that contribute toward the detection of depressed individuals.

Previous studies (Smirnova et al., 2018; Newell et al., 2018) have shown that the language of depressed individuals is riddled with features such as excessive usage of first person pronouns and negatively valenced words. Factors such as length of text, sentence type (single-clause vs multi-clause) and sentence style(generally ruminative prose) are also indicative of depression. Therefore we avoid arbitrary resampling and introducing summarization components to our framework as it not only adds to the amount of time taken to run the model but also because losing representations of such crucial information would inevitably lead to skewed results. Instead, a clustering and ranking approach is adopted to pick out the user's most relevant posts. This results in fewer posts per user and saves both training time and memory consumption. This also allows us to retain crucial user information and highlights common linguistic characteristics made by the user. The first step is to compute the embeddings for all posts made by each user using a pre-trained transformer model. For our experiments, post level representations are generated using BERT (Devlin et al., 2019) as a sentence encoder. We then use our clustering and ranking algorithm to find the subset of posts that would be a good representation of all the posts made by the user.

Various subset sizes were experimented with using both k-means and means clustering. Our experiments show that K-means gives a more diversified range of topics than the mean representations. However, in certain instances, some posts nearest to the centre of the clusters are duplicates. This is because the nearest posts from the centre of two or more clusters were identical. Therefore, the mean clustering method is preferred as it only considers unique posts near the centre of the embeddings. While the resulting representation of topics is not diverse in nature, this allows us to single out users who are prone to ruminative thinking and frequently exhibit other signs of depression as well.

Mean clustering: First, compute the mean of all the posts of a particular user. The top 'n' posts nearest to the mean by Euclidean distance are selected. Here, 'n' is the desired subset size of posts that would represent the user. Then compute the median representation for each user. This essentially returns the posts that are most similar to all the other posts made by the user. Thus, we eliminate a large number of posts in favour of the ones that most closely represent the majority.

4.2. Domain-Specific Pre-training

The final clustered posts of each user are tagged with a user ID and each post is then taken as input to the post-level RoBERTa (Liu et al., 2019) classifier. However, models such as RoBERTa are not trained to generate representations tuned to a specific domain. We therefore use 39248 unlabelled entries from our Mental Health Blog dataset to perform a variation of semi-supervised training. Our model tailors a pre-trained RoBERTa model to obtain more effective representations for our model. This is done by feeding posts from our Mental Health Blog dataset into a RoBERTa language model and taking the weights of all the layers except the final dense layer. This is then used to initialize the RSDD user-based classification model instead of training weights from scratch with random initialization. By tuning the RoBERTa parameters using its masked language modeling and sentence prediction tasks, we are able to effectively add domain-specific elements that represent depression on social media. These results also outperform popular models such as word-embeddings+LSTM architectures and are known to improve model performance significantly (Gururangan et al., 2020).

4.3. Tackling Imbalance

The ratio of positive to negative samples in the RSDD training data is 1:10.57. The imbalanced data often generates frustrating and disruptive results due to the bias that classifiers have towards the majority class. The number of majority labels overpowers the minority, and as a result, the features of the minority class are treated as noise. This leads to a high probability of misclassification of the minority class compared to the majority class. Resampling methods are very popular for dealing with this. Oversampling techniques such as SMOTE (Synthetic Minority Oversampling Technique) only increases the amount of memory usage and time. While under-sampling by clustering the abundant class is more feasible, we aim to address the issue of imbalance without tampering with the amount of data.

Since the dataset is skewed in nature, using a regular cross-entropy loss function gives us high recall and low precision. To correct this, class weights can be used to penalize the misclassification made by the minority class. A weighted cross-entropy loss is implemented such that the assigned class weights are inversely proportional to their respective frequencies. This consequently sets a higher class weight for the minority class and reduces the weight of the majority class.

4.4. Majority Voting Approach

Our final domain-specific RoBERTa model outputs labels for each post made by the user. However, it is difficult to determine the number of depression posts (or the degree of depression that must be represented in these posts) required to assign a final label to the user.

The most straightforward approach would be to simply take the most occurring label as the final assigned user label. This is called majority voting. Given that the posts we are considering represent the user's 'n' most relevant posts, we postulate that the 'degree' of depression in the user can be ascertained by simply taking the mode of the outputted 'n' post labels. However, our experiments show that although majority voting produces decent results at a user level, it performs poorly at a post level (resulting in an F1 score of around 65%). This is because while feeding training inputs to the classifier, all posts of the user are binarily labelled as strictly depressed or non-depressed. But there may be a significant number of posts among the top 'n' posts written by a depressed user that discuss a variety of topics unrelated to depression. Not all posts made by depression-tagged users were actually posts about depression.

An alternative to majority voting would be to concatenate posts into a single vector representing each user and use a confidence score to rule out posts with less confidence. The model is then retrained after removing these instances. However, there are caveats to this approach. It would require us to repeat this process on multiple models and take multiple runs to get an unbiased result before removing the least confident posts. This would only increase the amount of training time, which is against our objective. The batches/posts may also be too long to process with limited compute resources. Semi-supervised approaches could be considered, in addition to our pretrained RoBERTa model, to generate a weak label for the depressed user's posts and increase accuracy. However, as our dataset is already skewed in nature, generating weak labels to supplement our architecture would only serve to skew our results further. Hence, architectures such as Hierarchical Attention Networks are preferred to enhance our post to user-level classification accuracy.

4.5. Hierarchical Attention Network (HAN)

As our task can be perceived as a document (or a cluster of sentences) classification, we exploit its hierarchical structure (i.e., words form sentences and sentences come together to form a document). In this approach, representations of the chosen cluster of sentences are aggregated into a document representation. Because words and sentences in different documents can have different meanings (or word senses) depending on context (e.g., homonyms), the same word or sentence may be more (or less) notable in a different context. For this, two layers of attention are added at the word and sentence levels (Yang et al., 2016). This allows the model to pay more or less attention to individual words and sentences in a context-specific manner when constructing the representation of the clustered sentences. The class-weighted, domain-specific transformer model is used to encode the words of each post.

Word encoder For word representations, the entire sentence is passed to the transformer model. Mean pooling of the fifth to second last hidden states is taken and the vector corresponding to the word's position is extracted. Annotations for a given word are obtained by concatenating the forward and backward hidden states of a bidirectional GRU.

Word Attention The attention mechanism is used to extract the words that are important to the meaning of the sentence and the representation of such informative words is aggregated to form a sentence vector.

Sentence Encoder and Sentence attention Similar to word encoder and word attention, the same steps are followed to get the representation of a user from the sentence vectors obtained in the last step.

5. Experiments

We evaluate the proposed methods against several baselines and provide analyses. Our code is available here[2].

5.1. Evaluation Metrics

We adopt several standard metrics for classification such as Accuracy(represented as 'Acc'), Precision('P'), Recall ('R') and the F1 score('F1'). Our implementation utilizes libraries such as Pytorch (Paszke et al., 2019), HuggingFace (Wolf et al., 2020) and Scikit-learn (Pedregosa et al., 2011).

5.2. Baselines

Previous work that has been done on the RSDD dataset has largely cut down on the size by randomly sampling sections of the dataset, imposing parameters, and preprocessing based on the number of users, maximum number of posts per user, and maximum length of the posts.

- **Transformers:** We study the performance of popular transformer-based models, namely BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019). Observations were made using the default parameters to predict a given post.
- **X-A-BiLSTM**: This model (Cong et al., 2018) uses an attention-BiLSTM to enhance classification capacity. It also uses XGBoosting, a popular mechanism for targetting the data imbalance problem, as it offers a way to tune the training algorithm to pay more attention to misclassification of the minority class using the scale_pos_weight hyperparameter.
- **User-CNN**: In this user-level model(Yates et al., 2017), each post is fed to a convolutional neural network, which then performs average pooling. The resulting post representations are then merged with a second convolutional layer to create a user representation. The loss function used was categorical cross-entropy. Their results are generated by arbitrarily sampling posts before using them as input. Their CNN-R model randomly sampled 1500 posts per user, while their CNN-E samples the earliest 400 posts made by the user. In contrast, our architecture utilizes the entire dataset and beats the previous models in terms of both performance and efficiency (as can be seen in Table 9).

5.3. Experimental settings

Before clustering, we adopt the REDUCE_MEAN pooling strategy to get the sentence embeddings from bert-as-a-service (Xiao, 2018), which takes the average of the fifth to second last hidden states on the time axis. Our hardware configuration includes 10x Intel Xeon CPUs, 1x GeForce RTX 2080 TI graphics card, and 30GBs of RAM. For training, validation, and testing sets, we have used the default splits provided to us by the RSDD authors. Each split contains approximately 3,000 diagnosed users and 35,000 non-depressed users. We perform preprocessing by lowercasing text and removing certain non-alpha-numeric characters, stopwords[3], and extra spaces. The data is then tokenized using HuggingFace's tokenizer. HuggingFace is also used to implement RoBERTa for finetuning, and our architectures using transformers models and hierarchical attention networks were implemented in Pytorch (Paszke et al., 2019). All our experiments were run with a batch size of 64. For the hierarchical attention model, we set the word embedding dimension as 200 and the GRU dimension as 50. We also use stochastic gradient descent with 0.9 momentum, and the learning rate is picked by performing grid search on the validation set. At the end of the hierarchical attention network, we get the user embeddings which are passed through dense hidden layer, and softmax, which serves as the binary classifier head.

[2] `https://bit.ly/325udru`

[3] `https://www.nltk.org/`

5.4. Experimental Results

5.4.1. Clustering representations: Varying cluster parameters

Table 3 shows the results for the top ’n’ clustered posts per user. The number of posts in our clustering framework was experimentally decided for ’n’, and the distance from the mean is calculated using Euclidean distance. Several values of ’n’ were tested, including 5, 10, 15, 20, and 50 posts per user. While the n=50 subset performs the best, the difference in scores is consistently increasing less with the increase in ’n’ values. The accuracy seems much higher than the F1 scores because the model is not learning much and is predicting the majority label well.

n	Acc	P	R	F1	Time/epoch
5	86.75	**0.92**	0.2	0.32	7 min
10	89.54	0.89	0.26	0.40	16 min
20	90.1	0.9	0.31	0.46	30 min
50	**90.5**	0.91	**0.33**	**0.48**	71 min

Table 3: Results on top ’n’ clustered posts per user.

5.4.2. Dealing with Imbalance

Our class weighted cross-entropy loss function (results shown in Table 4) penalizes misclassification of the minority class and helps in reducing the effect of imbalance. We see our results improve drastically compared to Table 3, which uses the regular cross-entropy loss function. Once again, we see that the 50 sentence clustered subset consistently performs the best, which shows that more data per user leads to better results.

n	Acc	P	R	F1	Time/epoch
5	75.9	0.57	0.68	0.56	7 min
10	74.31	0.57	0.69	0.56	15 min
20	75.67	0.57	0.70	0.57	30 min
50	**78.78**	**0.58**	**0.70**	**0.59**	70 min

Table 4: Results with Class weighted Cross-Entropy Loss

5.4.3. Adding Domain-specific Elements

We find a significant improvement with each clustered subset by initializing the weights of the RSDD model using weights from our self-trained mental health language model, as shown in Table 5. We also compare the amount of time taken per epoch (in minutes) with not just prominent clusters but also with the fully unclustered dataset(represented as ’Total’). Though we get the best results with the entire dataset, we believe that the difference in scores with the n=50 model does not warrant the amount of time needed for using the fully unclustered dataset.

n	Acc	P	R	F1	Time/epoch
5	70.24	0.68	0.56	0.61	7min
10	78.61	0.72	0.61	0.66	15min
20	80.06	0.76	0.62	0.68	29min
50	82.2	0.77	0.64	0.69	69min
Total	**82.7**	**0.79**	**0.64**	**0.70**	1343min

Table 5: Results for the domain-specific RoBERTa model with class weighting

5.4.4. Final Diagnosis (Post → User level classification)

Experimenting with majority voting (Table 6) and Hierarchical Attention Architectures (Table 7) shows that the HAN model pays special attention to specific words and sentences and allows us to label a user as depressed more accurately. In contrast, the majority voting framework disregards nuances in misclassified posts before assigning a final label to the user. Because performing majority voting does not require running another model, the user-level results are near-instantaneous, and therefore, this takes the same amount of time as the domain-specific post-level classification model in Table 5.

n	Acc	P	R	F1	Time/epoch
5	77.82	0.64	0.58	0.61	7 min
10	82.57	0.66	0.61	0.63	15 min
20	81.99	0.66	0.63	0.65	29 min
50	**83.3**	**0.67**	**0.63**	**0.65**	69 min

Table 6: Majority Voting (MV) Results

n	Acc	P	R	F1	Time/epoch
5	76.41	0.7	0.59	0.64	23 min
10	81.36	0.77	0.6	0.67	42 min
20	81.39	0.77	0.61	0.68	96 min
50	**84.8**	**0.79**	**0.66**	**0.72**	180 min

Table 7: Hierarchical Attention Network (HAN) Results

5.5. Final Analysis

From our experimental results, we can clearly see how the addition of each subsystem influences the framework and increases the model’s overall performance.

Although the F1 scores increase with an increase in ’n’ value, the difference between F1 scores across all experiments decreases as ’n’ increases. Therefore, applying the model with extremely high ’n’ values cannot be afforded if we were to extend our scope to running such systems on real-time social media.

Since the RSDD dataset is highly skewed in nature, we note high precision and poor recall in Table 3, especially when taking a fewer number of ranked posts. This results in bad macro F1 scores. Replacing the default function with our class-weighted cross-entropy loss, we can see the increase in overall F1-score in Table 4 as the balanced class weights try to normalize the skewed effect. The addition of domain-specific elements by finetuning RSDD on our pretrained mental

Model	Acc	P	R	F1	Time/epoch
Post-level, Mean Clustering	**90.5**	**0.91**	0.33	0.48	71 min
Post-level, Clustered+Class-weighted	78.78	0.58	**0.70**	0.59	70 min
Post-level, Clustered+Class-weighted+Domain-specific	82.2	0.77	0.64	0.69	69min
Post-level, Unclustered+Class-weighted+Domain-specific	82.7	0.79	0.64	**0.70**	1343min
User-level, Majority Voting	83.3	0.67	0.63	0.65	69 min
User-level, HAN	**84.8**	**0.79**	**0.66**	**0.72**	180 min

Table 8: Ablation Study

Model	P	R	F1
BERT	0.54	0.51	0.52
XLNet	0.55	0.57	0.55
RoBERTa	0.55	0.56	0.56
X-A-BiLSTM	0.69	0.53	0.60
UserCNN-E	0.59	0.45	0.51
UserCNN-R	0.75	0.57	0.65
Domain-specific, MV	0.67	0.63	0.65
Domain-specific, HAN	**0.79**	**0.66**	**0.72**

Table 9: Comparing baseline results

health RoBERTa model (Table 5) shows a significant improvement of 11 points in F1 score from the best performing subset. For assigning the user label, we can see the effect of hierarchical attention models in Table 7, which gives an increase of 7 F1-score over majority voting. Though it takes longer to run, this compromise is justified because of this increase in performance, the significant reduction in memory load, and because the time taken is much less than it would be if we were to run any of the baselines on the unclustered dataset.

Our results also showcase the efficiency of our models as we can distinctly view the time vs. performance tradeoff. We can see in Table 5 that the models with clustered sentences take a lot less time than the full model does, without compromising a lot in terms of the quality of results. Taking a subset of n=50 posts per user gives near similar results while simultaneously taking $1/19^{th}$ of the time it would take to train on the entire dataset.

Considering the performance of our baselines in Table 9, we can see that our model shows significant improvement over the basic transformer models, as well as over previous work done on the dataset. The poor performance of such popular transformer models confirms the complexity of our problem. Our domain-specific HAN model outperforms the best-performing transformer model by 16 macro-F1 points. We also notice that the XA-BiLSTM and CNN-R models are the best performing baseline models as they tackle issues of imbalance and size of the dataset in some form. However, both variants of our proposed user-level framework (i.e., majority voted and hierarchical attention models) outperform all baselines and are efficient in their modelling as well. Our best model is the domain-specific HAN model, as it considers the context of words in each post before assigning a final label to the user. The addition of our mental health dataset to the framework has also proved to result in the highest increase in performance amongst all the subsystems.

6. Conclusions

We propose and compare methods to efficiently classify depressed users in social media. We also provide the Mental Health Blog dataset, which contributes integral domain-specific information while simultaneously emulating the degree of informality with which such topics are discussed on social media. Leveraging this dataset has proved to improve the performance of our depression detection mechanism. Our depression detection system also considers various challenges one faces when factoring in the size and inherent imbalance of topics on the web. Experimental results demonstrate how each subsystem contributes to the overall performance of our framework. Our clustering mechanism ensures we focus on the most relevant features of the user without arbitrarily losing vital information, while our hierarchical deep learning network effectively picks out important words or sentences from classified posts and ultimately allows us to label a user as depressed. This also helps in reducing the load on compute resources and reduces time constraints. Therefore, we are capable of running our system on large datasets, which can eventually be extended to include the entirety of social media. This would allow us to reach out to those in need of assistance, promote healthy lives, and ensure the social well-being of all on an Internet that is otherwise known for its toxicity.

7. Future Work

Our ultimate goal is to be able to forecast and reach out to depressed social media users to provide them with help and resources before their condition escalates to suicidal ideation. The work presented in this paper is our first step towards achieving this and identifying such users across the vast expanse of social media. In the future, we aim to develop language generation models to interact with these individuals and provide them with the necessary help. Adding images and other multimedia content to the scope could provide more contextual knowledge to not just depression detection models but also for diagnostic and dialogue systems.

8. Bibliographical References

Asad, N., Pranto, M., Afreen, S., and Islam, M. M. (2019). Depression detection by analyzing social media posts of user. pages 13–17, 11.

Beck, A. (1983). *Cognitive therapy of depression: New perspectives*. New York: Raven Press.

Burgess, E. (2019). Collaborative self-management of depression. pages 38–42, 11.

Choudhury, M. D. and De, S. (2014). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*.

Choudhury, M. D., Counts, S., Horvitz, E., and Hoff, A. (2014). Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work amp; Social Computing*, CSCW '14, page 626–638, New York, NY, USA. Association for Computing Machinery.

Cong, Q., Feng, Z., Li, F., Xiang, Y., Rao, G., and Tao, C. (2018). X-a-bilstm: a deep learning approach for depression detection in imbalanced data. *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1624–1627.

Coppersmith, G. A., Dredze, M., Harman, C., Hollingshead, K., and Mitchell, M. (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. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Fersko, H. (2018). Is social media bad for teens' mental health? *UNICEF*, Oct.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks.

Homan, C. M., Lu, N., Tu, X., Lytle, M. C., and Silenzio, V. M. (2014). Social structure and depression in trevorspace. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work amp; Social Computing*, CSCW '14, page 615–625, New York, NY, USA. Association for Computing Machinery.

Karim, F., Oyewande, A. A., Abdalla, L. F., Chaudhry Ehsanullah, R., and Khan, S. (2020). Social media use and its connection to mental health: A systematic review. *Cureus*, 12(6):e8627, June.

Li, G., Zhou, X., Lu, T., Yang, J., and Gu, N. (2016). Sunforum: Understanding depression in a chinese online community. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work amp; Social Computing*, CSCW '16, page 515–526, New York, NY, USA. Association for Computing Machinery.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Losada, D. E. and Crestani, F. A. (2016). A test collection for research on depression and language use. In *CLEF*.

Manikonda, L. and Choudhury, M. D., (2017). *Modeling and Understanding Visual Attributes of Mental Health Disclosures in Social Media*, page 170–181. Association for Computing Machinery, New York, NY, USA.

Newell, E. E., McCoy, S. K., Newman, M. L., Wellman, J. D., and Gardner, S. K. (2018). You sound so down: Capturing depressed affect through depressed language. *Journal of Language and Social Psychology*, 37(4):451–474.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E. Z., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Louppe, G., Prettenhofer, P., Weiss, R., Weiss, R. J., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830.

Pennebaker, J. W., Boyd, R. L., Jordan, K. N., and Blackburn, K. G. (2015). The development and psychometric properties of liwc2015.

Pyszczynski, T., Holt, K., and Greenberg, J. (1987). Depression, self-focused attention, and expectancies for positive and negative future life events for self and others. *Journal of personality and social psychology*, 52:994–1001, 06.

Richards, C. S. and OHara, M. W. (2014). *The Oxford Handbook of Depression and Comorbidity*. Oxford University Press.

Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Chua, T.-S., and Zhu, W. (2017). Depression detection via harvesting social media: A multimodal dictionary learning solution. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3838–3844.

Shen, T., Jia, J., Shen, G., Feng, F., He, X., Luan, H., Tang, J., Tiropanis, T., Chua, T.-S., and Hall, W. (2018). Cross-domain depression detection via harvesting social media. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 1611–1617. International Joint Conferences on Artificial Intelligence Organization, 7.

Smirnova, D., Cumming, P., Sloeva, E., Kuvshinova, N., Romanov, D. V., and Nosachev, G. (2018). Language patterns discriminate mild depression from normal sadness and euthymic state. *Frontiers in Psychiatry*, 9.

Wainberg, M. L., Scorza, P., Shultz, J. M., Helpman, L., Mootz, J. J., Johnson, K. A., Neria, Y., Bradford, J.-M. E., Oquendo, M. A., and Arbuckle, M. R. (2017). Challenges and opportunities in global mental health: A research-to-practice perspective. *Curr. Psychiatry Rep.*, 19(5):28, May.

(2017). "depression: let's talk" says who, as depression tops list of causes of ill health, Mar.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Xiao, H. (2018). bert-as-service. `https://github.com/hanxiao/bert-as-service`.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. H. (2016). Hierarchical attention networks for document classification. In *NAACL*, pages 1480–1489. Association for Computational Linguistics, 01.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Yates, A., Cohan, A., and Goharian, N. (2017). Depression and self-harm risk assessment in online forums. pages 2968–2978, 01.

Zogan, H., Razzak, I., Jameel, S., and Xu, G. (2021). Depressionnet: Learning multi-modalities with user post summarization for depression detection on social media. SIGIR '21, page 133–142, New York, NY, USA. Association for Computing Machinery.