Early detection of mental disorder via social media posts using deep learning models

Amanda Sun

Princeton High School

[a]151 Moore St

Abstract

Mental health, which has as equally important effects on people's life as physical health, is receiving more and more attention nowadays, especially with a significant increase of pressure brought by the fast-paced evolution of technology and society. The diagnosis of mental health symptoms, however, mostly relies on the interpretation of languages and behaviors by experienced psychologists, who are not accessible for the great population. Depression causes cognitive and motor changes that affect speech production: reduction in verbal activity productivity, prosodic speech irregularities, and monotonous speech have all been shown to be symptomatic of depression. In this study, we aim to provide a deep learning based model that could give an initial diagnosis of mental health problems for individuals and screen the risk of developing mental health issues. This AI-driven model focuses on the understanding and analysis of people's daily public comments/posts and captures the peoples' mental health status embedded in the semantic and syntactic structure in those online posts.

Keyword

Mental health, depression detection, deep learning, artificial intelligence

1. Introduction

Individual mental health problems have received more and more attention nowadays as it threatens the well-being of people, as much as physical health problems. According to Mental Health America (MHA)'s report in 2019 [1], nearly 20% of adults, which is equivalent to nearly 50 million Americans, experiencing mental illness and suicidal ideation continues to increase, with nearly 5% of adults having serious thoughts of suicide. Over half of the adults with mental illness remain untreated. Meanwhile, the situation for the youth is not

optimistic either, with over 15% of youth experienced severe depression and 60% of those who do not receive any mental health treatment. The lack of treatment is partially due to the scarcity of psychotherapy resources, but more so caused by the unawareness of the existence of the mental illness and a lack of timely warnings.

Unlike most physical illnesses, there is no standard medical test to diagnose mental illness. Instead, feelings, symptoms, and behaviors are usually used by psychologists and diagnosis heavily relies on the psychologists' interpretation of those observations. Social media provides platforms for people to express their feelings via posts in text and those text data embed abundant information about people's feelings and emotions. It was reported in [2] that people with depression are more likely to post tweets with negative emotional sentiments. And many efforts have been conducted to understand the risk of mental illness via social platform data. Feature engineering is vastly performed to generate attributes such as linguistic styles and social engagement to build traditional machine learning classifiers (e.g., support vector machine) for mental disorder prediction [3, 4, 5].

Hand engineering features is not an effective way to understand the information embedded in the posts as those features may not capture every detail in the raw text. With the significant amount of text data on social media, it is possible to build end-to-end natural language processing (NLP) models to enable the early detection of mental disorders. Both convolutional neural network (CNN) and long short term memory network (LSTM) have been applied to user content from social media [6, 7, 8]. The challenge associated with those end-to-end models is the requirement of a large amount of labeled data for training to guarantee satisfying prediction accuracy.

Language model pre-training is effective for improving performances on dif-

ferent natural language processing (NLP) tasks, including the mental disorder detection via social media posts in this study. There are several large-scale pre-trained language models developed that provide useful language embeddings that can be directly used for downstream language tasks (e.g., sentiment classification, text generation, question answering, etc). Embeddings from Language Model, aka, ELMo [9], utilizes a bidirectional language model to learn contextualized word representation in an unsupervised learning setup. The ELMo model is task-specific and needs to be trained separately for different tasks. A generative pre-training transformer (GPT) [10] was then proposed based on a transformer decoder architecture and supervised fine-tuning process for downstream task applications. GPT model architecture can be easily generalized to different NLP tasks with the pre-trained model and the general GPT framework was shown to provide superior performance than existing models back in time when it was proposed. Later GPT-2 [11] and GPT-3 [12] were published as the upgraded versions, with each version containing a 10x larger number of parameters than the previous version and therefore, more powerful and yet more computationally expensive to use.

Bidirectional Encoder Representations from Transformers (BERT) [13] utilizes a similar idea as the GPT model, which trains a large language model and is fine-tuned on specific tasks with the generic model architecture. BERT utilizes a multi-layer bidirectional transformer encoder and is trained on two tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP). MLM randomly masks 15% of the words/tokens in the sequence and the model tries to predict the masked words using other context words in the sequence. NSP is motivated by the observation that many NLP tasks, such as question answering, involve the understanding of the relationships between neighboring sentences and attempts to learn if the second sentence is the next sentence in

the data.

The pre-training language models discussed above have not been extensively studied and applied for mental disorder detection. In this study, we investigate several methods, which consist of deep learning networks and the state-of-the-art pre-training language model for mental illness detection via social media posts. We want to explore and validate if natural language processing models can find subtle language patterns based on the raw text of online posts to detect mental health issues, and if the detection performance could be improved by more state-of-the-art language models.

2. Data Collection and Preprocessing

There is a lot of textual data flooded to social media because of the increase of social media usage that has given researchers the opportunity to try to examine the emotions from the text. These data may aid in the analysis of feelings and provide valuable insight into the sudden discrepancy in the user's personality traits as reflected in one's posts. The goal of this work is real-time mental health status classification on online tweets, Reddit posts, comments, etc using the developed natural language processing model. The model performance will heavily rely on the data quality and therefore, the collection and preprocessing of data with/without mental health issues is one of the most important contributions of this study.

Despite the recent research on identifying different mental health problems via text data, the public datasets in this area are rather limited. Meanwhile, it is challenging to get tweets and Reddit posts that indicate depression with manual labeling, which requires tremendous labor work. Instead, we take the advantage of existing public datasets, as well as the Twitter data filtering tool for the data collection. Existing datasets that we use here include **TalkLife** [14]

and **Dreaddit** [15]. **TalkLife** is the largest global peer-to-peer mental health support network. It enables seekers to have textual interactions with peer supporters through conversational threads. The dataset contains 6.4M threads and 18M interactions (seeker post, response post pairs). **Dreaddit** is a new text corpus of lengthy multi-domain social media data for the identification of stress, which we only extract the text data and corresponding labels.

Besides reformulating existing datasets into our need, we also collect data from Twitter and Reddit directly that indicate depression, loneliness, anxiety, etc. These datasets were collected so that an AI agent can be trained to evaluate mental health status via verbal expression. We developed a script using Twint (an open-source data scraper provided by Twitter to scrape data on its platform) that filters out data based on hashtags such as '#depressed', '#depression', '#hopeless' and so on. Meanwhile, we use Reddit PRAW API to collect data from subReddits that are likely to have mental health issues discussed, including interpersonal conflicts and mental illness, which was summarized by Sharma et al. [16] with 55 mental health-focused subReddits. The scraped data is then manually reviewed to remove noisy data samples.

The above discussion describes how the negative samples that indicate mental health issues are collected. When a machine learning classifier is trained, positive samples that indicate no mental health problems are equally important as negative samples. We randomly sample a subset (8000 samples) from the sentiment140 [17] dataset as positive samples. The sentiment140 is a public dataset that contains 1.6 million tweets. Combining the positive and negative samples collected via methods described above, we arrive at our dataset shown in Table. 1

Table 1: dataset

labels	number of samples
mental health issue	7952
no mental health issue	8000

The dataset collected above contains raw text, emojis, URLs, etc. To better assist the machine learning model training, proper pre-preprocessing is required. The following data preprocessing steps are performed:

- Convert all cases in text to lower cases
- Remove symbols such as emojis and URLs
- Expand contractions (e.g., can't → can not)
- Remove punctuation and stop words

Those preprocessing steps can effectively reduce the data noise and facilitate the training of the machine learning model.

We then analyze the word frequency in posts indicating mental disorders and visualize the results in Fig. 1. The larger the font size is, the more frequently the word appears in the negative posts with mental health issues. It is observable that keywords such as 'depression' and 'anxiety' make up a large proportion in the negative posts.

3. Methods

Five models including one baseline model are trained and evaluated over the dataset we created.

3.1. Random predictor model

We use the random predictor model as a natural baseline model, where the label distributions in the training data are memorized and then used to

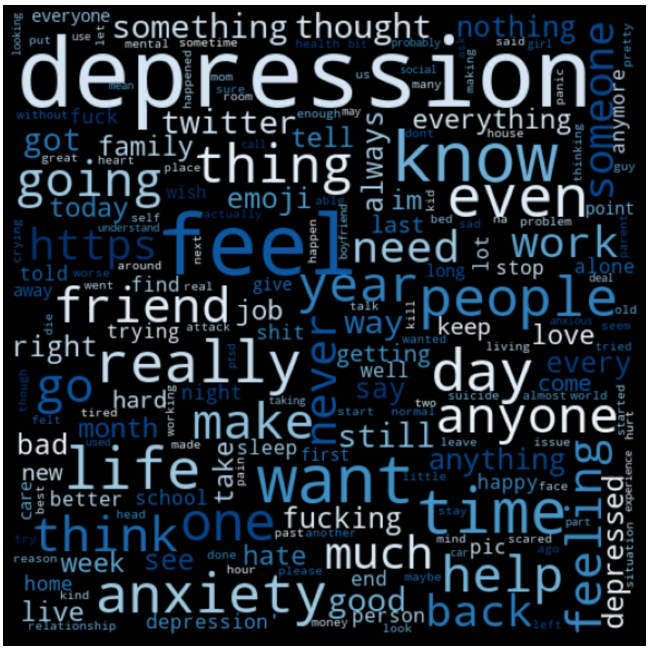

Figure 1: Word cloud visualization for posts indicating mental health problems. The larger the word font size is, the more frequently it appears in the posts/tweets with mental health problems.

predict a random class conforming to the label distribution of each new test case regardless of features we have for the test case.

3.2. sentence2vec

We implement a 100-dimension *sentence2vec* model, which returns the average of the word embeddings in the input sequence (or sentence, noting that one sample may contain multiple sentences simultaneously). The pre-trained word embeddings provided by *gensim* [18], an open-source natural language processing package, are obtained by applying GloVe on the Twitter dataset. We specifically use the 'glove-Twitter-100' version of the pre-trained embeddings.

3.3. Deep learning LSTM model

An LSTM model with an embedding layer was implemented to take the tokenized sequence as input and predict the mental health status directly. The

architecture of the LSTM model is shown in Table 2.

Table 2: LSTM model architecture

Layer	Output Size	# Parameters
Embedding Layer	(None, 105, 128)	256000
Spatial Dropout 1D	(None, 105, 128)	0
LSTM	(None, 196)	254800
Dense	(None, 2)	394

The 105 in the output size from Table 2 denotes the maximum number of tokens within the training data. All training samples are padded to match that number so the model can batch process the data. All embedding weights in our LSTM model are randomly initialized and no pre-trained weights are utilized. The LSTM model is trained end-to-end using the training data we collected.

3.4. pre-trained BERT

BERT pre-trains language representations to build practical models that can be widely used for different tasks. We first use pre-trained BERT to extract high-quality language features, then train a logistic regression model over the extracted language features to classify the processed tweets/posts as either postive or negative. The extracted features for each processed raw text is a vector of size 768, which is an embedding for the tweet/post that we can use for classification.

In this study, instead of using the vanilla BERT model, we use DistilBERT [19], which is a smaller, faster, and lighter version open-sourced by HuggingFace. It has only 60% size of the vanilla BERT but runs 60% faster and matches 97% language understanding performance.

Several data processing steps are required before the processed data in Section 2 is fed to DistilBERT:

- Tokenization – break words into tokens, add [CLS] and [SEP] tokens to

the start and end of the sentence respectively, and substitute tokens with corresponding ids

- Padding – pad all lists of tokens to the same size to represent the input as a 2D array for batch processing
- Masking – create a mask array that has the same size as the input array and masks the padding so that the BERT model won't be confused by the padded area.

Three embeddings are combined as input embedding in BERT: Token embedding, segment embedding and position embedding, as illustrated in Figure 2. Wordpiece tokenization embedding is used in the BERT model so that unusual words can be split into sub-word units. Sentence embeddings are designed to differentiate two sentences, and position embeddings are learned to reflect the position of words in the sequence. Trained BERT-based models are rarely used as-is, but are generally fine-tuned (transfer training) on the target dataset. The pre-trained BERT model grasps the semantic and syntactic relationships between words from a large dataset. Because of the shared semantic and syntactic information in different language datasets, a new model can be simply fine-tuned on a small target dataset to obtain superior performance and avoid overfitting.

3.5. Fine-tuned BERT

In Section 3.4, the pre-trained DistilBERT model is used as a feature extractor purely and all weights in the model are frozen. Because the pre-trained weights carry much language information, it allows fine-tuning on specific tasks with a much smaller dataset and less training time. In this subsection, we have an extra fully connected layer added after the general BERT model and fine-tune the overall model over the processed dataset.

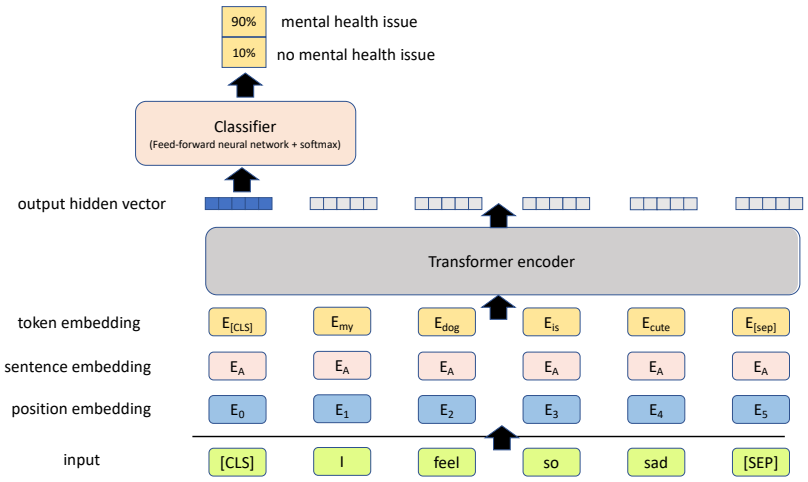

Figure 2: BERT structure overview.

4. Experiments

Our experiments sought to understand if the proposed NLP models can detect the mental health status purely based on textual languages such as tweets and Reddit posts and the corresponding accuracies.

4.1. Quantitative Evaluations

The quantitative testing results by different models described in Section 3 are provided in Table. 3. Because both the training and testing data are pretty balanced, accuracy and macro-F1 score (as well as other metrics such as precision and recall) will be similar. Therefore, we only use accuracy as the evaluation metric for simplicity. The test accuracies from different language models (e.g., sentence2vec, LSTM, pre-trained model, fine-tuned model) are significantly higher than the random model, which shows that NLP models could find language statistics and patterns to help detect mental health status abnormalities.

The BERT model (pre-trained and fine-tuned) outperforms other language models such as sentence2vec and LSTM, which shows that more powerful language model architectures can improve the performance of detecting tweets/posts that reflect mental health issues. The fine-tuned BERT model is able to achieve above 96% classification accuracy over the random test data, which demonstrates the effectiveness of fine-tuning over the pre-trained language models.

Table 3: Quantitative results

Model	Accuracy
Random	0.511
Sentence2vec + LR	0.857
LSTM	0.934
pre-trained BERT + LR	0.947
Fine-tuned BERT	0.963

4.2. Qualitative Analysis

We now show resulting examples with fine-tuned BERT model in Table 4, where the first column shows the raw texts (without any processing) from different posts, the second column indicates the true label of the posts, with 0 representing normal posts and 1 denoting posts with depression or mental health problems, the third column shows the prediction by our best model, i.e., the fine-tuned BERT.

The first example, 'Just checked my user timeline on my blackberry ...', is a pretty normal post, as we can safely conclude that no mental disorders can be seen from that post. And our language model makes the correct prediction. The second example, 'Just heard gun shots in my neighborhood!!!', however, is very interesting as the fine-tuned BERT model classifies this post as potentially showing signs of a mental disorder while the label shows it is a normal post. The label here could be viewed as error or data noise because it is undeniable that the post reflects a certain level of anxiety and fear, which may result in

potential mental disorder. The fact that the proposed language model is able to detect it demonstrates the robustness of the model. Similarly, the third and fourth example also proves that the proposed language model is more than a simple 'keyword detector', as very few words in those examples exist in the high-frequency set shown in Fig. 1.

Table 4: Qualitative analysis for examples

Posts	Label	Prediction
Just checked my user timeline on my blackberry, it looks like the twanking is still happening Are ppl still having probs w/ BGs and UIDs?	0	0
Just heard gun shots in my neighborhood!!!	0	1
If anyone will listen. I'm in a bad place right now. I could really use a friend.	1	1
It cleared up and I was okay but. On Monday I was thinking about humans and how the brain works and it tripped me out I got worried that because I was thinking about how the brain works that I would lose sleep and I did. That night was bad just like last time. Also yesterday my sleep was bad I woke up like every hour of the night just like last time.	1	1

5. Conclusions and Future Work

The application of different natural language processing models on the early detection of mental health disorders via social media posts is explored and we found that deep learning language models help detect potential mental disorders based solely on raw text in social media. Among all the models investigated, the advanced fine-tuned BERT model proves to be the most effective model for detecting mental disorder signs from the text. Compared with the baseline model where sentence embeddings are derived by averaging word embeddings, The LSTM, the pre-trained BERT, and the fine-tuned BERT lead to obvious better classification performance.

For future work, we want to explore if the performance of fine-tuned BERT model could be further improved with a larger training dataset. Also, the lan-

guage model investigated in this study could be powered online to provide real-time analysis of individual mental health status by analyzing comments as well as about public mental health statistics. Moreover, the attention mechanism visualization could be investigated and applied to explain the model behaviors, i.e., why the model classifies specific posts/tweets as ones indicating mental health problems.

References

[1] M. Reinert, D. Fritze, T. Nguyen, The state of mental health in america 2022.

[2] G. Park, H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, M. Kosinski, D. J. Stillwell, L. H. Ungar, M. E. Seligman, Automatic personality assessment through social media language., Journal of personality and social psychology 108 (6) (2015) 934.

[3] M. De Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting depression via social media, in: Seventh international AAAI conference on weblogs and social media, 2013.

[4] A. G. Reece, A. J. Reagan, K. L. Lix, P. S. Dodds, C. M. Danforth, E. J. Langer, Forecasting the onset and course of mental illness with twitter data, Scientific reports 7 (1) (2017) 1–11.

[5] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, H. Ohsaki, Recognizing depression from twitter activity, in: Proceedings of the 33rd annual ACM conference on human factors in computing systems, 2015, pp. 3187–3196.

[6] G. Gkotsis, A. Oellrich, S. Velupillai, M. Liakata, T. J. Hubbard, R. J.

Dobson, R. Dutta, Characterisation of mental health conditions in social media using informed deep learning, Scientific reports 7 (1) (2017) 1–11.

[7] J. Du, Y. Zhang, J. Luo, Y. Jia, Q. Wei, C. Tao, H. Xu, Extracting psychiatric stressors for suicide from social media using deep learning, BMC medical informatics and decision making 18 (2) (2018) 77–87.

[8] J. Kim, J. Lee, E. Park, J. Han, A deep learning model for detecting mental illness from user content on social media, Scientific reports 10 (1) (2020) 1–6.

[9] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proc. of NAACL, 2018.

[10] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training.

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (8) (2019) 9.

[12] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, arXiv preprint arXiv:2005.14165.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805.

[14] K. Saha, A. Sharma, Causal factors of effective psychosocial outcomes in online mental health communities, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 14, 2020, pp. 590–601.

[15] A. Sharma, A. S. Miner, D. C. Atkins, T. Althoff, A computational approach to understanding empathy expressed in text-based mental health support, in: EMNLP, 2020.

[16] E. Sharma, M. De Choudhury, Mental health support and its relationship to linguistic accommodation in online communities, in: Proceedings of the 2018 CHI conference on human factors in computing systems, 2018, pp. 1–13.

[17] N. Friedrich, T. D. Bowman, W. G. Stock, S. Haustein, Adapting sentiment analysis for tweets linking to scientific papers, arXiv preprint arXiv:1507.01967.

[18] R. Rehurek, P. Sojka, Gensim–python framework for vector space modelling, NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic 3 (2).

[19] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108.

Acknowledgement

Dr. Tang provided instructions on literature review (papers to start with) and study materials (youtube videos) to learn natural language processing models. I read representative papers in NLP and implemented the model, conducted the experiments (including hyperparameter tuning) and analyzed the model results. Then I also iterated over improving the model performance. In the end, I organized the data analysis and wrote the report under the guidance of Dr. Tang.