Detection of Depression and Anxiety on Social Media

Hillary Ngai
20609183

Abstract

Anxiety and depression are often unreported, undetected, and therefore, untreated due to their internalizing symptoms and the social stigma associated with mental illness [1]. The lack of proper treatment can cause symptoms to worsen, potentially leading to substance abuse and suicide [2]. Therefore, early detection of depression and anxiety is critical for rapid intervention to reduce the escalation of such disorders [2]. Nowadays, many individuals turn to social media forums for mental health support [2]. Comments and posts on such online social networks provide insight into how people self-disclose and discuss mental illness. Using a dataset of scraped Reddit comments, this project aims to classify anxiety and depression in comments. Natural language processing, machine learning methods, and neural network architectures were used to build, tune, and evaluate models that detect anxiety and depression in social media text entries. Two baseline models were built using tf-idf—Logistic Regression and SVM—and three transformer-based language representation models were also developed—a BERT-based model, a RoBERTa-based model, and an XLNet-based model. Ultimately, the XLNet model achieved the best performance yielding a 92.8% accuracy and 92.4% F1 score.

Introduction

Major depressive disorder (MDD) and generalized anxiety disorder (GAD) is among the most prevalent psychiatric disorders, affecting more than 300 million people globally [1]. Depression and anxiety are both internalizing disorders making it difficult to detect such medical conditions [1]. Furthermore, mental illness is a taboo subject almost everywhere in the world [1]. This causes many cases of depression and anxiety to go unnoticed, unreported, and therefore, untreated [2]. Early detection of depression and anxiety is critical for rapid intervention, which can potentially reduce the escalation of the disorders [2].

Individuals suffering from mental health issues often turn to anonymous social media forums for mental health support [3]. Comments and posts on such online social networks provide insight into how people self-disclose and discuss mental illness, such as depression and anxiety [3]. Unlike other social media platforms (e.g. Facebook and Instagram), Reddit is a social news sharing site where registered members are not associated with permanent online identities [3]. This unique platform allows Reddit users to freely make posts or comments disclosing sensitive information while also remaining anonymous [3].

Depression and anxiety influence how individuals use language [3]. Therefore, mental illness forums on Reddit provide valuable linguistic information for the detection of depression and anxiety in social media text entries. The automated detection of depression and anxiety in social media text entries can aid in identifying at-risk individuals through large-scale, passive monitoring of social media [4]. This automation can compliment other technologies to help tease out actionable insights, track the progress of treatment, and advance accessible mental illness medical care such as therapy chatbots [4].

This project aims to use natural language processing, machine learning methods, and neural network architectures to build, tune, and evaluate models that classify Reddit text comments as “depressed”, “anxious” or “not depressed and not anxious”.

Background Review

Various research has been done to study the relationship between mental health and linguistics to provide insight into depression and anxiety detection. Sigmund Freud, an Austrian neurologist, wrote about Freudian slips to reveal subconscious thoughts and feelings of the speaker [5]. One paper writes about the excessive use of words that convey negative emotions, such as “lonely”, “sad”, or “miserable”, amongst individuals suffering from depression and anxiety [6]. Such subjects were also found to use significantly more first-person singular pronouns and less second-person and third-person pronouns [6]. This finding suggests that subjects diagnosed with depression and anxiety are more focused on themselves and less connected with others [6]. Such lexical patterns provide valuable insight in mental illness diagnostics [6].

The rapid and vast adoption of social media provides a rich and abundant source of textual data and social metadata to capture the behavioural tendencies of users [7]. Natural language processing and machine learning have been used to perform sentiment analysis of social media posts [7]. For example, previous research has involved building models to identify depression in Reddit posts [7]. The paper uses Term-Frequency/Inverse-Document Frequency (TF-IDF) weighted combinations of word n-grams and Linguistic Inquiry and Word Count (LIWC) features to train a Linear Support Vector Machine (SVM) [7]. This model achieved 81.8% accuracy in the task of classifying a Reddit post as “depressed” or “non-depressed” [7].

Language representation models with transformer-based architecture have significantly advanced the field of natural language processing yielding state-of-the-art results in a wide variety of NLP tasks, including question-and-answering (e.g. SQuAD v2.0) and natural language inference tasks (e.g. MultiNLI) [9]. State-of-the-art language representation models include Bidirectional Encoder Representations from Transformers (BERT) and its derivatives—A Robustly Optimized BERT Pre-training Approach (RoBERTa) and Generalized Autoregressive Pre-training for Language Understanding (XLNet) [8] [9] [10]. These general-purpose language representation models are special because they are unsupervised and deeply bidirectional [8]. The models are trained using an enormous amount of unlabelled text corpus, such as Wikipedia. In contrast to previous word embedding models which read text sequences from left to right, transformer-based models represent text sequences based on the previous context and next context—making it bidirectional [8]. This is achieved using a technique called Masked LM (MLM) which randomly masks out a percentage of the words in the input, runs the entire sequence through a deep bidirectional transformer encoder, and then predicts on the masked words [8]. BERT, RoBERTa, and XLNet differ slightly in their pre-training method. BERT uses Masked ML (MLM) and Next Sentence Prediction (NSP), RoBERTA uses Masked LM (MLM), and XLNet uses permutation-based modeling [8] [9] [10]. Pre-trained versions of BERT, RoBERTa, and XLNet are made publicly available for fine-tuning on downstream NLP tasks or feature extraction via transfer learning for text classification [10].

One paper used transfer learning with BERT to perform fine-grained sentiment classification with five different classes—very negative, negative, neutral, positive, and very positive [11]. Transfer learning with BERT achieved 94% accuracy outperforming Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Bidirectional LSTM, and Convolutional Neural Network (CNN) model architectures [11].

Application/Dataset

Reddit is broken up into more than a million communities known as subreddits, each of which covers a different topic of discussion [1]. Reddit users can anonymously post and comment in these subreddits [1]. Every submission can be replied to and upvoted or downvoted by other users [1].

The dataset was created by scraping the newest and most upvoted posts from three subreddits: /r/depression, /r/Anxiety, and /r/AskReddit. The Python library, PRAW, was used to scrape each subreddits. The dataset contained 3741 Reddit comments in total, consisting of 1172 comments from the subreddit /r/depression, 1175 comments from the subreddit /r/anxiety, and 1394 comments from the subreddit /r/AskReddit. The dataset is relatively balanced with an approximately equal number of samples in each class. Each comment was pre-processed to remove non-ASCII characters, tabs, newlines, and user mentions. Stop words were not removed from the text since recent work suggests depression is correlated with the frequent use of first-person pronouns [6]. The dataset was then concatenated, randomly shuffled, and split into 80%-20% ratio for the training and testing sets. The final dataset contained a column of text comments and a column of labels. Each comment was labelled with “depressed”, “anxious”, or “not anxious and not depressed” based on the subreddit the comment was posted on.

We chose these three subreddits because /r/ depression is described as a “a supportive space for anyone struggling with depression”, /r/Anxiety is a discussion and support for “suffers and loves ones of any anxiety disorder”, and /r/AskReddit is similar to the other two subreddits in that it is also of question-and-answer format, except the questions can be of any topic [1]. The following table contains example texts of each class.

Table 1: Sample texts of each class

Text	Label
Does anyone else feel normal one minute then wish you were dead the next?	Depressed
I’m having panic attacks because of quarantine. I think its time for me to try medications instead of talking. How does it work, during coronavirus? When 2 years ago I hit a new low in regards of panic attacks I was feeling strong heartbeats, difficulties breathing, dizziness in general. I had to live with them for about a month, 24/7.	Anxious
Would you watch a show where a billionaire CEO has to go an entire month on their lowest paid employee’s salary, without access to any other resources than that of the employee? What do you think would happen?	Not depressed and not anxious

Proposed Scheme/Algorithms

Five different machine learning models were built to classify Reddit comments as “depressed”, “anxious”, or “not depressed and not anxious”.

Two baseline models were proposed—Logistic Regression and Linear Support Vector Machine (Linear SVM). Comments were first converted to a matrix of term frequency-inverse document frequency (tf-idf) features to serve as inputs for these baseline models. Tf-idf weights measure how statistically important each word is to a document in a corpus [1]. Each row in the tfi-idf matrix was regularized using L2 normalization and terms that appear in less than two comments were ignored. The Logistic Regression model was configured by setting the inverse of regularization strength to one. The SVM model used a linear kernel and the penalty parameter of the error term was configured to one.

Three different pre-trained transformer-based language representation models—BERT-base-uncased, RoBERTa-base, and XLNet-base-cased—were fine-tuned on the downstream task of detecting depression and anxiety in Reddit comments. All three pre-trained models are case-sensitive and trained on the English language. We developed three different models using the above mentioned pre-trained transformer-based language representation models. The models were developed using the following architecture.

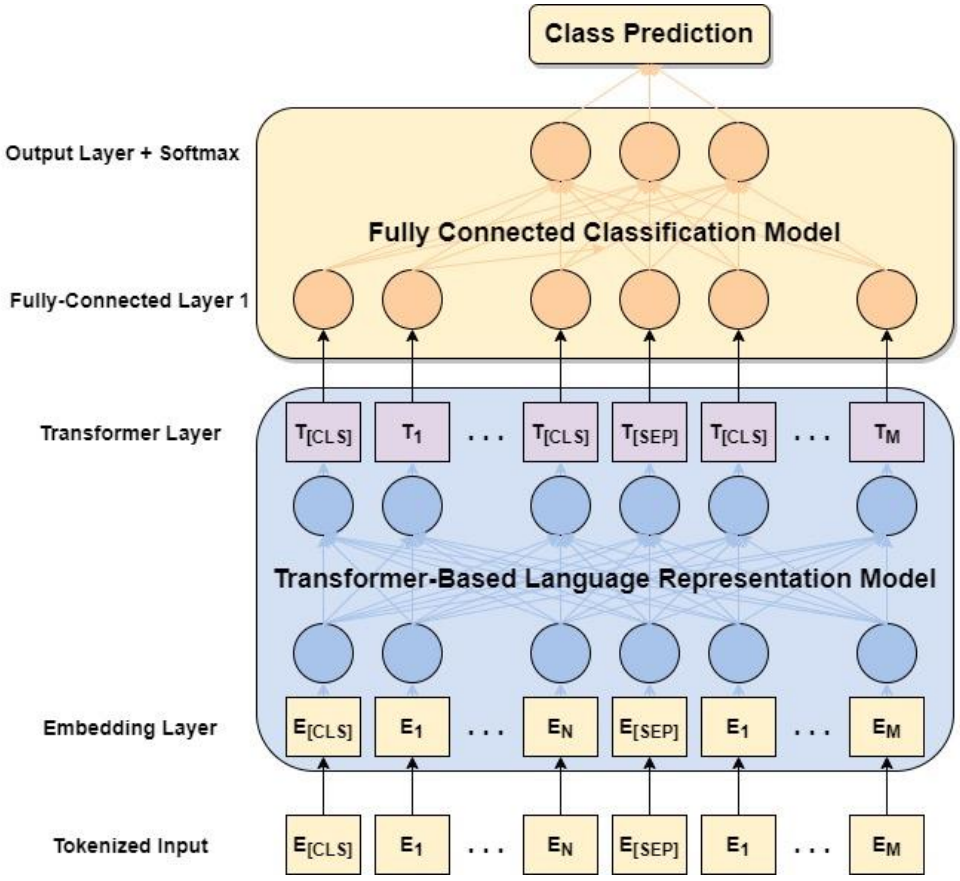

Figure 1: Transformer-based language representation model architecture for classifying Reddit comments as “depressed”, “anxious”, or “not depressed and not anxious”

Each model was trained with a batch size of 32 for 10 epochs using the Adam optimizer with a learning rate of $2e^{-5}$ and an epsilon value of $1e^{-8}$. Note that 5% of the training data was used for validation. The parameters of the transformer-based language representation models are shown in the table below.

Table 2: Parameters of transformer-based language representation models

		Transformer-Based Language Representation Model		
		BERT-base-cased	**RoBERTa-base**	**XLNet-base-cased**
Parameter	**Number of Transformer Layers**	12	12	12
	Number of Hidden Units	768	768	768
	Number of Self-Attention Heads	12	12	12
	Total Trainable Parameters	110M	125M	110M
	Pre-training Data	3,300M words from BookCorpus and English Wikipedia	33,000M words from BookCorpus, English Wikipedia, CC-News, OpenWebText, and Stories	3,300M words from BookCorpus and English Wikipedia
	Pre-training Data Size	16GB	160GB	16GB
	Dropout Probability	0.1	0.1	0.1
	Activation Function	Gaussian Error Linear Units (GELU)	Tanh	Gaussian Error Linear Units (GELU)

The rest of the pre-training parameters are mentioned in the papers of each of the language representation models [9] [10] [11].

The fully connected classification component is composed of three hidden layers with 512, 256, and 128 neurons from the first hidden layer to the third hidden layer. The output layer and the Softmax layer each have three neurons corresponding to “depressed”, “anxious”, and “not depressed and not anxious”. Each neuron in the fully connected layers uses the Rectified Linear Unit (ReLU) activation function.

Experiments and Results

The performance metrics used to validate the proposed models are shown below.

$$Accuracy = \frac{\sum_{class}(TruePositive_{class} + TrueNegative_{class})}{Total\ Number\ of\ Samples} \quad (1)$$

$$Macro\ Precision = \frac{1}{Class_{num}} \sum_{class} \frac{TruePositive_{class}}{TruePositive_{class} + FalsePositive_{class}} \quad (2)$$

$$Macro\ Recall = \frac{1}{Class_{num}} \sum_{class} \frac{TruePositive_{class}}{TruePositive_{class} + FalseNegative_{class}} \quad (3)$$

$$Macro\ F1\ Score = \frac{1}{1/Recall_{macro} + 1/Precision_{macro}} \quad (4)$$

Where $Class_{num}$ is the number of classes. The quantitative results of the models are summarized in the table below.

Table 3: Quantitative results of the models

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression with tf-idf	0.781	0.769	0.769	0.769
Linear SVM with tf-idf	0.793	0.783	0.783	0.783
BERT-base-cased	0.919	0.915	0.914	0.914
RoBERTa-base	0.922	0.920	0.920	0.920
XLNet-base-cased	0.928	0.924	0.924	0.924

In theory, the baseline accuracy for this classification task is approximately 33.33%—there is approximately a one in three chance of classifying a comment as "depressed", "anxious", or “not depressed and not anxious” as the training set was roughly evenly split between classes. All models developed performed better than this theoretical baseline.

As seen in Table 3, the transformer-based language representation models performed better than the baseline models achieving accuracies ranging from 91.9% to 92.8% and F1 scores ranging from 91.4% to 92.4%. This corresponds to an improvement in accuracy ranging from 12.6% to 14.7% and an improvement in F1 scores ranging from 13.1% to 15.5%. This improvement in performance is because tf-idf weights only captures the frequency of words in documents relative to the frequency of a word in the corpus. Transformer-based language representation models, however, can learn representations of words based on the left and right context in a sequence of text [8]. For example, the word “bank” can have different meanings depending on the bidirectional context of the text sequence. “I need to go to the bank to deposit money” has a different context than “I am going to the river bank to skip stones”. This is because BERT is a transformer-based language representation model which is designed to pre-train deep bidirectional representations from unlabelled text [8]. Another advantage of the transformer-based language models is that the models are pre-trained on an enormous text corpus allowing the models to learn general patterns in language [8]. This knowledge is captured in

contextualized word embeddings which can be extracted as features via transfer learning for supervised NLP tasks [8]. This explains why the BERT, RoBERTa, and XLNet models only required 10 epochs to achieve superior performance.

The RoBERTa model yielded slightly better results than the BERT model achieving a 92.2% accuracy and 92.0% F1 score—0.3% better accuracy and 0.6% better F1 score than the BERT model. This improvement in performance is likely because RoBERTa is built upon BERT's architecture except the Next Sentence Prediction objective is removed leading to better downstream task performance [9]. RoBERTa is also pre-trained on 10 times more data than BERT, for a longer time period.

Ultimately, the XLNet model performed the best yielding a 92.8% accuracy and 92.4% F1 score. This is a slight improvement in performance than the RoBERTa model—0.6% better accuracy and 0.4% better F1 score than the RoBERTa model. This improvement in performance is because BERT ignores the dependencies between mask positions in the Masked Language Modeling (i.e. the fill-in-the-blank task) causing a pre-train-finetune discrepancy [10]. XLNet uses a generalized autoregressive pre-training method that learns bidirectional contexts by maximizing the log likelihood of all possible factorization sequences and overcomes the limitations of BERT with auto-regression [10]. This generalized autoregressive pre-training method generally leads to better downstream NLP task performance—in this case, classifying Reddit comments as "depressed", "anxious", or "not depressed and not anxious".

One issue with the custom dataset is that the most popular and newest /r/depression, /r/anxiety, and /r/AskReddit comments were immediately classified as "depressed", "anxious", and "not depressed and not anxious", respectively. However, this may not always be the case since it is possible that a comment in /r/depression or /r/anxiety could be more lighthearted and therefore, not entirely representative of depression. To improve this dataset, each comment could have been manually labelled by a psychologist or comments of users who have claimed to have been diagnosed with depression or anxiety could have been labelled based on their diagnosis.

Furthermore, it is possible for a comment to be both "anxious" and "depressed". For example, one of the comments misclassified as "anxious" by the XLNet model when the true label was "depressed" was, "Everything hurts. My head hurts, my bones hurt, my arms and legs hurt, my eyes hurt. I had a massive panic attack last night with my S.O. next to me, not knowing what the heck to do." Perhaps the task should have been set up as a multi-label classification task where "depressed" and "anxious" may both be assigned to each comment, instead of a multiclass classification task.

Another potential improvement is to explore a character-based CNN approach since social media text can be unique to the user and exhibit an informal writing style. For instance, social media users tend to use emojis and are prone to grammatical errors. The transformer-based language representation models were pre-trained on formal corpus, such as Wikipedia, which can have a dissimilar writing style to that of social media text.

Detecting depression and anxiety using only social media comments is a nontrivial task. Overall, the XLNet model achieved impressive results yielding a 92.8% accuracy and 92.4% F1 score.

References

[1] F. Cacheda, D. Fernandez, F.J Novoa, and V. Carneiro, “Early Detection of Depression: Social Network Analysis and Random Forest Techniques,” J Med Internet Res, 2019.

[2] Statistics Canada, “Health at a Glance,” Mental and substance use disorders in Canada, 27-Nov-2015. [Online]. Available: https://www150.statcan.gc.ca/n1/pub/82-624-x/2013001/article/11855-eng.htm. [Accessed: 18-Nov-2019].

[3] M. Trotzek, S. Koitka, and C. M. Friedrich, “Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 3, pp. 588–601, Jan. 2020.

[4] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, “Detecting depression and mental illness on social media: an integrative review,” *Current Opinion in Behavioral Sciences*, vol. 18, pp. 43–49, 2017.

[5] A. T. Beck, Depression: Clinical, Experim., Theoretical Aspects. Philadelphia, PA, USA: Univ. Pennsylvania Press, 1967.

[6] Y. R. Tausczik and J. W. Pennebaker, “The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods,” *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, Aug. 2009.

[7] J.T. Wolohan, M. Hiraga, and A. Mukherjee, “Detecting Linguistic Traces of Depression in Topic-Restricted Text: Attending to Self-Stigmatized Depression with NLP,” Proceedings of the First International Workshop on Language Cognition and Computational Models, pages 11–21, 2018.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019.

[9] Yang, Zhilin, Dai, Yang, Yiming, Carbonell, Jaime, Salakhutdinov, Ruslan, and Q. V., “XLNet: Generalized Autoregressive Pretraining for Language Understanding,” Conference on Neural Information Processing Systems (NeurIPS), 2020.

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” International Conference on Learning Representations (ICLR), 2020.

[11] M. Munikar, S. Shakya and A. Shrestha, "Fine-grained Sentiment Classification using BERT," Artificial Intelligence for Transforming Business and Society (AITB), Kathmandu, Nepal, 2019, pp. 1-5.