Artificial Intelligence and Social Computing, Vol. 28, 2022, 155–163
https://doi.org/10.54941/ahfe1001458

Detecting Potential Depressed Users in Twitter Using a Fine-Tuned DistilBERT Model

Miguel Antonio Adarlo and Marlene De Leon

Department of Information Systems and Computer Science, Ateneo de Manila University, Loyola Heights, Quezon City, Philippines

ABSTRACT

With the spread of Major Depressive Disorder, otherwise known simply as depression, around the world, various efforts have been made to combat it and to potentially reach out to those suffering from it. Part of those efforts includes the use of technology, such as machine learning models, to screen a potential person for depression through various means, including social media narratives, such as tweets from Twitter. Hence, this study aims to evaluate how well a pre-trained DistilBERT, a transformer model for natural language processing that was fine-tuned on a set of tweets coming from depressed and non-depressed users, can detect potential users in Twitter as having depression. Two models were built using the same procedure of preprocessing, splitting, tokenizing, training, fine-tuning, and optimizing. Both the Base Model (trained on CLPsych 2015 Dataset) and the Mixed Model (trained on the CLPsych 2015 Dataset and a half of the dataset of scraped tweets) could detect potential users in Twitter for depression more than half of the time by demonstrating an Area under the Receiver Operating Curve (AUC) score of 65% and 63%, respectively, when evaluated using the test dataset. These models performed comparably in identifying potential depressed users in Twitter given that there was no significant difference in their AUC scores when subjected to a z-test at 95% confidence interval and 0.05 level of significance ($p = 0.21$). These results suggest DistilBERT, when fine-tuned, may be used to detect potential users in Twitter for depression.

Keywords: Twitter, Depression, Transformer Models, DistilBERT, Health and Well-being

INTRODUCTION

Major Depressive Disorder, simply known as depression, is a mental health condition that affects a significant portion of the global population. Various efforts have been made to address the burden of depression by potentially reaching out to those suffering from it. However, the utilization of mental health programs remains low due to the inaccessibility of mental health services and the social stigma attached to seeking mental help.

A possible way to reach people afflicted with depression is through social media, wherein their state of mind can be captured from their posts as narratives that show their innermost selves and state of mental health. Using data

obtained from social media, various predictive models have been developed trying to address issues of mental health, ranging from rule-based models detecting sentiment in text, machine learning models addressing potential at-risk individuals, to even multi-task neural network models classifying a variety of mental health issues. One architecture that shows promise would be the use of models based on the Transformer architecture for mental health tasks. However, the Transformer architecture and its models, such as DistilBERT, have not been used in widespread applications for mental health due to their novelty. Hence, this study aims to determine how well a pre-trained DistilBERT that was fine-tuned on a set of tweets coming from depressed and non-depressed users can detect potential users in Twitter as having depression.

RELATED WORK

The Transformer deep learning model architecture was conceived in 2017 to utilize the concept of attention within speech for improved natural language processing (NLP) (Vaswani *et al.*, 2017). From the original architecture, various modifications were made to fit many different needs, such as next sentence prediction and sentiment analysis, among others. BERT, or Bidirectional Encoder Representations from Transformers, was one such modification by making use of the encoder block from the original Transformer architecture and the bidirectional approach to parsing text to obtain context and meaning embedded within language. Its pre-trained language model version was shown to perform well across differing NLP tasks without much modification, aside from fine-tuning several parameters (Devlin *et al.*, 2019). From BERT came DistilBERT, a "smaller, faster, cheaper, and lighter" version of BERT. Where BERT has 100 million trainable parameters, DistilBERT has 66 million, and yet it still performs similarly to BERT across NLP tasks (Sanh *et al.*, 2019).

Within the field of using predictive models for mental health in social media, the Transformer architecture has not yet been employed widely. However, several studies have been carried out making use of Transformers in some fashion. A study by Wang et al. made use of BERT, among other models, to evaluate the potential risk for depression from posts made in the Chinese social media platform, Weibo. In this study, the BERT model performed the best, with a macro-F1 score of 0.538 from the four listed levels of risk from 0 to 3 (Wang *et al.*, 2019). Another study by Matero et al. used embeddings from BERT in a feature-based approach to model suicide risk on Twitter, with the generated embeddings improving the performance of the model (Matero *et al.*, 2019).

Applying the Transformer architecture along with offshoots like BERT and DistilBERT shows promise for detecting mental illnesses in social media by using the pre-trained language model and fine-tuning parameters for the task. However, additional studies, such as the use of DistilBERT in detecting potential users for depression based on their tweets, are still needed.

METHODOLOGY

The methodology for fine-tuning DistilBERT in detecting depression is divided into four parts, namely data collection, preprocessing, model creation, and model validation.

Data Collection

The first set of data came from the Computational Linguistics and Clinical Psychology (CLPsych) 2015 Shared Task. This dataset contained 1,711 Twitter users with 1,145 and 566 of them belonging to the training and test sets, respectively. Within the training set, there are 327 users with depression, 246 with Post-traumatic Stress Disorder (PTSD), and 572 as controls. As mental health may impact certain segments of the population more than others, each user identified as having depression or PTSD was matched with a control of the same age and gender. The controls were obtained from the Twitter API's Spritzer stream of data. Around 3,000 tweets from each user were included in the dataset. The tweets in this dataset were in the English language only (Coppersmith *et al.*, 2015).

This dataset was considered in training the models due to its usefulness in previous studies (Resnik *et al.*, 2015; Nadeem *et al.*, 2016; Amir *et al.*, 2017; Jamil *et al.*, 2017; Orabi *et al.*, 2018). However, data on PTSD were not included in this study as detecting this disorder was beyond its scope. Hence, Twitter users with PTSD and their matched controls were excluded from the dataset used in this study. Additionally, the test set was not used in training the transformer model as it lacks labels (Coppersmith *et al.*, 2015). Thus, the training dataset included 1,498,060 tweets, consisting of tweets from depressed individuals and tweets from their age and gender-matched controls.

The dataset used for testing and validating the Transformer models was a set of publicly available English tweets posted from 1 October 2020 to 1 December 2020. A previous study showed tracking tweets made within two months was enough given that longer periods tend to reduce the performance of the model. Queries were made for tweets that contained text identifying a diagnosis of depression, such as "I have been diagnosed with depression" (Coppersmith, Dredze, and Harman, 2014). Rather than using the Twitter API that can only collect the last 3,200 tweets or other scrapers that require a Twitter developer account, Twitter Intelligence Tool (TWINT) was used in this study. TWINT is a scraping tool built with Python that can scrape much more than the last 3,200 tweets. Aside from its ability to scrape certain keywords, users' likes, and followers, TWINT allows visualization of tweets and scraping based on language ("TWINT - Twitter Intelligence Tool," no date). Disingenuous statements or statements that do not refer to depression as a mental disorder were removed manually from the initially gathered tweets. Overall, tweets from 110 users, who declared a diagnosis of depression were obtained.

To find controls for these users with depression, 110 users were randomly selected from the Twitter API's Spritzer stream of data and their tweets for two months were scraped.

The two scraped datasets were then combined to form a dataset with labels for depressed and non-depressed users. Since this dataset can be contaminated with users mislabeled as not having depression, it was mainly used for validating the models' performance, though half of the collected users were used as part of the training set in building the Mixed Model.

Preprocessing

Preprocessing of the CLPsych 2015 and scraped datasets involved removing parts of the tweets that would make the data noisier (e.g., Twitter handles, hyperlinks, re-tweets, and non-alphanumeric characters, such as emojis). However, excluding words, such as pronouns and other stop words, was not carried out as these words could provide context on the emotional state of the user (Coppersmith *et al.*, 2016).

From all the gathered users and their tweets within both datasets, a sample of 100 tweets per user was then aggregated into a corpus of tweets. A sample of 100 tweets was chosen per user in this study due to limitations in tokenization, as DistilBERT only has a maximum of 512 token limit per text sequence. Aggregating all the tweets per user into a single corpus would end up being truncated automatically to fit within DistilBERT's sequence limit.

Model Creation

To create the Transformer model, the HuggingFace Transformer library and PyTorch library were used. The former allows for the use of pre-trained models in various tasks such as sentiment analysis, text classification, and question-answer pairs, among others, while the latter is a prerequisite to the former and aids in the transformation of data into tensors for use with the model, among other tasks (Paszke *et al.*, 2019; Wolf *et al.*, 2020). Pre-trained models have proven their generalizability for supervised learning tasks without having to build task-specific architectures (Devlin *et al.*, 2019). Thus, this study took a pre-trained model, in the form of DistilBERT, to train and fine-tune the training data from the CLPsych2015 dataset (Coppersmith *et al.*, 2015; Sanh *et al.*, 2019). Modifications, such as the addition of a pooling layer, were not done on the model to gauge the ability of the base DistilBERT in detecting depression in tweets. As such, the pre-trained model's structure for this study before fine-tuning is the same as shown in Sanh *et al* (2019).

The preprocessed CLPsych 2015 dataset wherein tweets were aggregated at the user level was split into a train-test split of 70%/30%, by way of scikit-learn's train-test-split method (Pedregosa *et al.*, 2011). The resulting split users were then tokenized and transformed for input into the transformer via the vocabulary used to pre-train DistilBERT. Afterward, the tokenized data was placed into DistilBERT for training and fine-tuning. The model was trained with a batch size of 16. In building the model, the Tune library was used to search for optimal hyperparameters, such as the number of epochs, the learning rate, and the weight decay (Liaw *et al.*, 2018).

The resulting embeddings from the trained and fine-tuned DistilBERT model were fed to a logistic regression layer for final labeling using the logistic

regression head that came with the classification task. The resulting fine-tuned model should be able to detect potential users in Twitter for depression. The built model is referred to as the Base Model.

The Mixed Model was built using the same procedure of splitting, tokenizing, training, fine-tuning, and optimizing. Its main difference from the Base Model is the data fed to train the model. Rather than just using the preprocessed data from CLPsych 2015, half of the users from the scraped dataset were also added to provide more data to train on.

The scikit-learn package was then used to generate the metrics for the models' performance (Pedregosa *et al.*, 2011). These metrics included accuracy, precision, recall, F1, and the Area under the Receiver Operating Curve (AUC).

Validation

The scraped dataset was used to validate the built transformer models in terms of how well these two models can predict the actual labels in the scraped dataset. The same dataset was also used for testing the generalizability of the built models.

Before this scraped dataset was inserted into the built model, it was converted into tokens via the tokenizer, the vocabulary used to pre-train DistilBERT. The sci-kit learn package was then used to obtain the performance of the Base and Mixed Models in terms of accuracy, precision, recall, F1, and AUC.

To compare the performance of the built models with reference to AUC, a z-test was done at a 95% confidence interval and 0.05 level of significance. This statistical test was carried out to show if there is a significant difference between the two models as regards their ability to predict labels on the remaining half of the users in the scraped dataset (i.e., data not used in training the second model).

RESULTS

This section shows the performance of the Base and Mixed Model in detecting potential users in Twitter for depression.

Base Model Performance

The Base Model's hyperparameters were 3 epochs, a learning rate of 3.39e-5, and a weight decay of 0.13. The resulting Base Model could correctly detect potential users in Twitter for depression at least 60% of the time as shown by its accuracy score of 64% and an F1 score of 66%. It could tell apart true positives from false positives to a degree with a precision score of 62%. The model could also discern true positives from false negatives 72% of the time, based on the recall score. Additionally, the Base Model could identify potential users in Twitter as having depression by 65% as shown by its AUC score (Table 1).

When evaluated using the full scraped dataset, the model's performance remained relatively stable, indicating that the Base Model trained on the preprocessed CLPsych 2015 dataset could be generalizable in identifying potential users in Twitter for depression based on their corpus of tweets

Table 1. Base model training metrics.

Epoch	Training Loss	Evaluation Loss	Accuracy	F1	Precision	Recall	AUC
1	0.68	0.66	0.59	0.63	0.56	0.73	0.59
2	0.60	0.68	0.63	0.69	0.59	0.83	0.63
3	0.52	0.67	0.64	0.66	0.62	0.62	0.65

Table 2. Base model evaluation metrics.

Evaluation Loss	Accuracy	F1	Precision	Recall	AUC
0.64	0.65	0.70	0.61	0.83	0.65

Table 3. Base model evaluation metrics – reduced evaluation dataset.

Evaluation Loss	Accuracy	F1	Precision	Recall	AUC
0.64	0.65	0.71	0.61	0.84	0.65

Table 4. Mixed model training metrics.

Epoch	Training Loss	Evaluation Loss	Accuracy	F1	Precision	Recall	AUC
1	0.68	0.66	0.61	0.54	0.60	0.50	0.60
2	0.65	0.65	0.63	0.49	0.70	0.37	0.62
3	0.53	0.63	0.66	0.58	0.69	0.50	0.65
4	0.39	0.66	0.67	0.61	0.69	0.54	0.67
5	0.27	0.72	0.65	0.61	0.63	0.60	0.64

(Table 2). The performance scores of the Base Model using the training and test sets of data were similar, with both sets able to correctly detect potential users in Twitter for depression 65% of the time as shown by its AUC. These metrics did not change meaningfully when evaluated using only the second half of the scraped dataset that was not used in training the Mixed Model (Table 3).

Mixed Model Performance

The Mixed Model's hyperparameters were 5 epochs, a learning rate of 4.19e-5, and a weight decay of 0.06. The built Mixed Model could correctly identify potential users in Twitter for depression at least 60% of the time as shown by its accuracy score of 65% and an F1 score of 61%. It could distinguish true positives from false positives 63% of the time based on its precision score. The model could also discern true positives from false negatives 60% of the time, based on the recall score. Furthermore, the computed AUC showed the model could identify potential depressed users in Twitter 64% of the time (Table 4).

When evaluated using the second half of users in the scraped dataset, the Mixed Model could detect potential users in Twitter for depression based on their tweets 63% of the time as shown by its accuracy and AUC scores. It could identify potential users in Twitter as having depression or not 66% of

Table 5. Mixed model evaluation metrics.

Evaluation Loss	Accuracy	F1	Precision	Recall	AUC
0.71	0.63	0.59	0.66	0.53	0.63

Table 6. Comparison of evaluation metrics between base and mixed models.

Model	Evaluation Loss	Accuracy	F1	Precision	Recall	AUC
Base	**0.64**	**0.65**	**0.71**	0.61	**0.84**	**0.65**
Mixed	0.71	0.63	0.59	**0.66**	0.53	0.63

the time based on its precision score. The reduced scores in F1 and recall, though, suggest that the Mixed Model had difficulty in telling apart true positives from false negatives (Table 5).

Comparison of Base and Mixed Models

Table 6 shows the comparison between the Base and Mixed Models in terms of their performance when tested against the second half of the scraped dataset. Apart from precision, where the Base Model performed worse than the Mixed Model, the Base Model had better metrics all-around, from a lower evaluation loss to higher accuracy, F1, recall, and AUC scores. Interestingly, the Base Model demonstrated a better recall score than the Mixed Model by 0.31. This result suggests that the Base Model could discern true positives from false negatives around 1.58 times better than the Mixed Model.

Using a z-test at 95% confidence interval and 0.05 level of significance, there was no significant difference in AUC scores between the Base and Mixed Models ($p = 0.21$). This result means that both models were comparable in terms of identifying potential users in Twitter for depression.

CONCLUSION AND RECOMMENDATIONS

Several metrics showed a Transformer model, such as DistilBERT, can be fine-tuned to detect potential depressed users on Twitter. Despite these promising results, the fine-tuned DistilBERT models in this study can still be improved. Future studies can do so by addressing some limitations of this research.

First, the data from the training set could be outdated already. More recent data should be used for training so that the dataset will be in line with modern discourse on Twitter.

Second, the scraped dataset used to test and validate the built models was not matched by age and gender due to data privacy concerns, among others. The training and test dataset should be similar and, if possible, both demographically controlled to ensure comparability. Gathering the test dataset with similar characteristics as the training set may improve the model's performance.

Third, the datasets may have contamination due to possible mislabeling of users. Future studies can benefit from having the datasets for training, testing,

and validation of the built models be evaluated and verified for mental illness by a clinical psychologist or psychiatrist.

Fourth, the hyperparameters used in fine-tuning the models' performance were not explored extensively due to constraints in hardware. Future studies can benefit from using more robust hardware for training and testing the built models.

Fifth, this study only used DistilBERT as a Transformer model to fine-tune. Other Transformer models can be considered as a base especially if more information can be kept within the changed parameters or architecture. Doing this may yield better results.

Sixth, the scope of this study only included fine-tuning a pre-existing model for a specific task, such as detecting potential depressed users based on their tweets. Further studies can also make use of the feature extraction method as previous research using BERT as a Transformer model suggests inputting contextualized word embeddings generated from BERT into another model can be helpful to produce better performance for a given NLP task (Devlin *et al.*, 2019). Other studies should also consider preprocessing further the training set before it is fed into the tokenizer and building more task-specific architectures that can support the Transformer model.

Lastly, this study focused on the detection of depression in tweets. Additional studies may opt to include the identification of other mental disorders in social media narratives aside from Twitter.

ACKNOWLEDGMENTS

The researchers would like to acknowledge Dr. Mark Dredze for providing the CLPsych 2015 Shared Task Data for use in this study, Ateneo de Manila University for offering opportunities to learn and work on this research study, and their family and friends for providing a wellspring of support in many things related and unrelated to this study, even in these tough times.

REFERENCES

Amir, S. *et al.* (2017) "Quantifying Mental Health from Social Media with Neural User Embeddings," in Doshi-Velez, F. et al. (eds) *Proceedings of the 2nd Machine Learning for Healthcare Conference*. Boston, Massachusetts: PMLR (Proceedings of Machine Learning Research), pp. 306–321. Available at: http://proceedings.mlr.press/v68/amir17a.html.

Coppersmith, G. *et al.* (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*. Association for Computational Linguistics. doi:10.3115/v1/w15-1204.

Coppersmith, G. *et al.* (2016) "Exploratory Analysis of Social Media Prior to a Suicide Attempt," in *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics. doi:10.18653/v1/w16-0311.

Coppersmith, G., Dredze, M. and Harman, C. (2014) "Quantifying Mental Health Signals in Twitter," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics. doi:10.3115/v1/w14-3207.

Devlin, J. *et al.* (2019) "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. doi:10.18653/v1/N19-1423.

Jamil, Z. *et al.* (2017) "Monitoring Tweets for Depression to Detect At-risk Users," in *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology - From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics. doi:10.18653/v1/w17-3104.

Liaw, R. *et al.* (2018) "Tune: A Research Platform for Distributed Model Selection and Training," *ArXiv*, abs/1807.05118.

Matero, M. *et al.* (2019) "Suicide Risk Assessment with Multi-level Dual-Context Language and," in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics. doi:10.18653/v1/w19-3005.

Nadeem, M. *et al.* (2016) "Identifying Depression on Twitter," *arXiv e-prints*, p. arXiv:1607.07384.

Orabi, A.H. *et al.* (2018) "Deep Learning for Depression Detection of Twitter Users," in *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. Association for Computational Linguistics. doi:10.18653/v1/w18-0609.

Paszke, A. *et al.* (2019) "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in Wallach, H. et al. (eds) *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., pp. 8024–8035. Available at: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Pedregosa, F. *et al.* (2011) "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, 12, pp. 2825–2830.

Resnik, P. *et al.* (2015) "Beyond LDA: Exploring Supervised Topic Modeling for Depression-Related Language in Twitter," in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics. doi:10.3115/v1/w15-1212.

Sanh, V. *et al.* (2019) "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," in *2019 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing for Embedded Applications (EMC2)*.

"TWINT - Twitter Intelligence Tool" (no date).

Vaswani, A. *et al.* (2017) "Attention is All You Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc. (NIPS'17), pp. 6000–6010.

Wang, X. *et al.* (2019) "Assessing depression risk in Chinese microblogs: a corpus and machine learning methods," in *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 1–5. doi:10.1109/ICHI.2019.8904506.

Wolf, T. *et al.* (2020) "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. Available at: https://www.aclweb.org/anthology/2020.emnlp-demos.6.