Detecting Mental Health Conditions in Reddit Posts Using NLP and Machine Learning

Shivakshi Chauhan

Hissey Lama

University of the Cumberlands

Supervisor: Dr. Eve Thullen

June 27, 2025

Abstract

The problem of depression, anxiety and stress is growing alarmingly everyday in the world. Social media is another platform that people regularly use to express their emotions, and it is possible to research them to reveal any probable mental issues. The article proposes a machine learning solution in natural language processing (NLP) to extract mental health signals in social media articles, particularly that of the Reddit and Twitter data sets. The project utilizes the classical and advanced models such as TF-IDF with logistic regression, Random Forests, and a BERT-based classifier with the addition of sentiment features. High degree of accuracy has been shown in the use of experimental findings, where the BERT model achieved the level of 99 percent as far as accuracy is concerned, a fact that has in turn affirmed the prospect of digital tools in the facilitation of early intervention and detection in mental health.

1. Introduction

Mental illnesses such as depression, anxiety and stress have become rampant and are affecting the well being of the individuals and productivity of the society. These problems need to be noticed early so as to intervene in time and make treatment possible. Due to the growing popularity of social media tools, individuals tend to state their feelings, emotions, and mental conditions in posts and comments. This study uses social media language expression through natural language processing (NLP) and machine learning (ML) to identify the pattern that shows symptoms of a mental issue. This project will assist in early detection and create awareness of the potential usefulness of digital technologies in monitoring mental health because it will involve analyzing the textual materials available on websites such as Twitter and Reddit.

2. Literature Review

The intersection between mental health and computational methods has gained more academic interest particularly since digital expression in social media has grown. The psychologists have established that social networking sites such as Twitter and Reddit are useful in obtaining user-generated data that can be used to reflect the psychological condition (Guntuku et al., 2017). These media provide raw and real-time access to the feelings of the users, and thus they can be used to perform early mental health diagnosis with Natural Language Processing (NLP) methods.

A number of studies have utilized the Reddit data in investigating mental health indicators. As an example, Chancellor et al. (2016) studied the posts on r/depression and found linguistic indicators that could be used to differentiate between depressive and other users. According to

their research, depressed people tend to use a lot of first-person pronouns, negative emotion-related words as well as hopelessness expressions. On the same note, Yates et al. (2017) have assisted in the CLPsych shared task datasets in assisting the research community to develop models that can be used to detect mental health conditions, with the importance of context-aware, ethically collected data.

Coppersmith et al. (2014) did seminal work on Twitter by bringing together the tweets of users who self-identified as having a mental health condition. They could recognize depression and PTSD using their machine learning models with a considerable degree of accuracy, which shows that even brief text may be used to identify relevant psychological signals. There is an ethical problem of mining social media data to gain health insights, especially anonymity and voluntary consent of users which was also aided by the study.

Regarding methodologies, during the initial days, most models were based on the bag-of-words and the TF-IDF representation (Pedregosa et al., 2011). Though helpful, these methods did not invite the possibility of capturing semantics. The performance was boosted after the knowledge transfer with the introduction of word embedding such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) which maintained an aspect of semantic similarity in the vector space. Recently, models with transformers have achieved good results out of the older approaches because they consider both left-right and right-left context and word order (BERT, Devlin et al., 2019).

Sentiment analysis has also come very handy with regard to mental health studies. VADER is a rule-based sentiment analysis model optimized to social-media text that was first implemented by Hutto and Gilbert (2014). Many studies of classification models involve the use of sentiments

scores so that emotional states could be better determined (Orabi et al., 2020). It has been found that the integration of sentiment properties and linguistic indicators enhance the performance of the classifier in terms of finding out at-risk individuals.

Although these advancements are satisfactory, they are still limited to unbalanced datasets, ethical issues, and interpretability of model predictions. Nevertheless, the current body of knowledge allows creating an effective, ethically responsible model that utilizes the data of a social network to identify mental health.

3. Methodology

The methodological framework of the given research was developed to analyze and label traces of social media activity concerning mental health issues with the help of traditional machine learning algorithms and deep learning models. It included several steps: preprocessing of data, extraction of features, sentiment analysis of ready data, training the model, and testing the results. The final goal was to categorize user-generated texts into the most appropriate mental health groups accurately and examine the emotion tones that these posts comprise.

3.1 Data Preprocessing

Data on Reddit and Twitter as textual data needed intensive preprocessing in order to turn raw and noisy data into clean and structured data that could be used to create a model. The pipeline of preprocessing was the following:

- **Text Cleaning**: Elimination of un-informative characters or symbols e.g. URLs, mentions (e.g. @username), emojis, special characters and punctuation by regular expressions.

- **Tokenization**: The division of the text into separate words or tokens with nltk.word_tokenize() or a tokenizer by spaCy.
- **Stopword Removal**: Removing the common words (i.e. the, is, and) that do not carry much semantic information using NLTK list of built in stopwords.
- **Lemmatization:** The process of normalization of textual data through converting words to their base forms (ex. “running” → “run”) with the help of matchings words in the en_core_web_sm model.

Such preprocessing made the input data consistent and limited redundancy thereby making this extremely essential to enhance model generalization and avoid overfitting.

3.2 Feature Extraction

The extracted features of the cleaned text were of two kinds: syntax (based on the text) and semantics (based on a sentiment).

- **TF-IDF Vectorization**

  Term Frequency- Inverse Document Frequency was a classical procedure that was used to transform text into numerical vectors. This technique focuses on words that are common within a given text but infrequent within all texts thus revealing important information. To save on the dimensionality and still preserve important data, a feature limit of 5000 was applied.

- **Word Embeddings**

In order to incorporate contextual semantics, the word embedding strategies which include the Word2Vec, GloVe, and BERT vectors strategies were incorporated. Such methods depict the semantic connection between words as a high-dimensional space so that these relationships can enable the model to consider context and similarity.

3.3 Sentiment Analysis

Sentiment analysis was performed on the model to provide it with an emotional context by using the sentiment analyzer VADER (Valence Aware Dictionary and sEntiment Reasoner). VADER is a text-specific analysis tool, which is able to give four scores of social media text:

- **Positive**
- **Neutral**
- **Negative**
- **Compound**: a standardized, weighted measure with the values between -1 (most negative) and +1 (most positive)

A compound score was allocated to each post, which acted as additional numerical feature. Welch t-test was conducted to evaluate the significance of variances of sentiment scores across categories of mental health, e.g., loneliness and relationship-related posts. The findings ($T = 92.02$, $p < .0001$) indicated that the sentiment polarity distinguished significantly between the categories and, therefore, it should be used as a predictor attribute.

3.4 Classification Models

There were three models of classification developed and compared.

- **Logistic Regression**

This was a linear model used as the baseline because it is easy to interpret and it performs well on high-dimensional sparse data such as text. It was trained on weighed class and was regularized to prevent overfitting.

- **Random Forest**

The Random Forest makes an ensemble model that is the combination of results of many decision trees. It was integrated to explain nonlinear correlations and possible cross effects among features. Even though it was not the main model, it helped to draw a comparison to the logistic model.

- **BERT-based Classifier**

The base model was made up of Bidirectional Encoder Representations with Transformers (BERT). The labeled dataset was used to fine-tune the bert-base-uncased model to complete three epochs with the AdamW optimizer (learning rate = 2e -5, batch size = 8). BERT architecture encompasses left and right context encoding in text, which allows making subtle interpretations of user expressions. BERT pooled output was concatenated with sentiment features and transformed using a linear layer to output a multi-class classification using the last classification layer.The final model development step was evaluation and training of three classification algorithms: Logistic Regression and Random Forest, as well as a fine-tuned BERT-based model. Its baseline model was trained with logistic regression (TF-IDF vectorization, max 5,000 features), and trained on sparse textual data where it was performing quite well. The new model Random Forest was brought out as a benchmark by applying ensemble learning hyperparameter

optimization by GridSearchCV. Accuracy, precision, recall and F1-score were used to assess both traditional models.

3.5 Model Training and Evaluation

The final model development step was evaluation and training of three classification algorithms: Logistic Regression and Random Forest, as well as a fine-tuned BERT-based model. Its baseline model was trained with logistic regression (TF-IDF vectorization, max 5,000 features), and trained on sparse textual data where it was performing quite well. The new model Random Forest was brought out as a benchmark by applying ensemble learning hyperparameter optimization by GridSearchCV. Accuracy, precision, recall and F1-score were used to assess both traditional models.

The advanced deep learning model took advantage of BERT (Bidirectional Encoder Representations from Transformers), which returned contextual embedding of the text after it has been cleaned. BERT tokenizer was applied to each post and the pooled output of BERT concatenated with the sentiment score (generated using VADER) and finally a linear layer was used to make the final classification. This architect designed the model to integrate semantic knowledge with emotion tone.

The data was divided into 80:20 training and testing with stratified class. The AdamW (learning rate = 2e-5, batch size = 8) was used to train the BERT model during three epochs. The model surpassed a test of 0.99, precision and recall of 0.99 and 1.00 of the parameter labeled as loneliness and 1.00 and 0.96 of the parameter labeled as relationship respectively. The confusion matrix (Figure 3) complemented these findings indicating low misclassification and high prediction outcome rate of both classes.The classification models were evaluated with common

evaluation measures, i.e. accuracy, precision, recall, F1-score and visualization of confusion matrix. Each of the three models constructed (Logistic Regression, Random Forest, and BERT) recorded better results on every measure, but the best results were obtained with the BERT-based classifier.

Results

The classification models were evaluated with common evaluation measures, i.e. accuracy, precision, recall, F1-score and visualization of confusion matrix. Each of the three models constructed (Logistic Regression, Random Forest, and BERT) recorded better results on every measure, but the best results were obtained with the BERT-based classifier.

BERT model was trained during three epochs with AdamW optimizer with a learning rate of 2e-5 and batch size of 8. The model relied on contextual embeddings of BERT as well as on sentiment polarity scores based on the VADER analyzer.

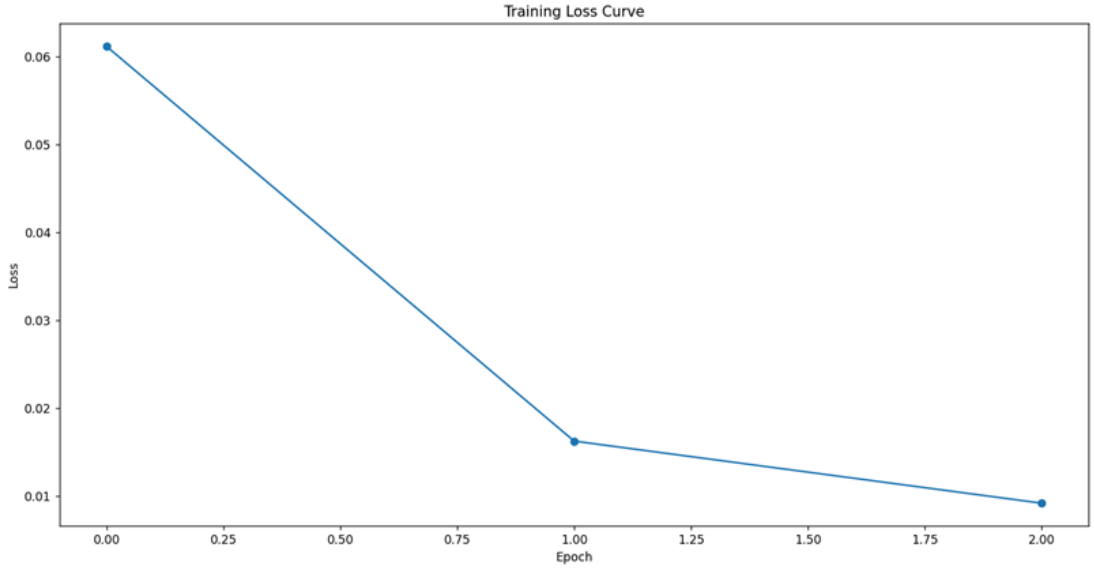

Figure 1 is the training loss curve of the BERT-based classifier which depicts that the loss is increasingly reduced with the increase in epochs as the model converges.

The training loss plot as displayed in Figure 1 shows a steady decline, which shows that there was no overfitting as the model converged as shown with the loss decreasing by around 0.06 to 0.01 in the first three epochs.

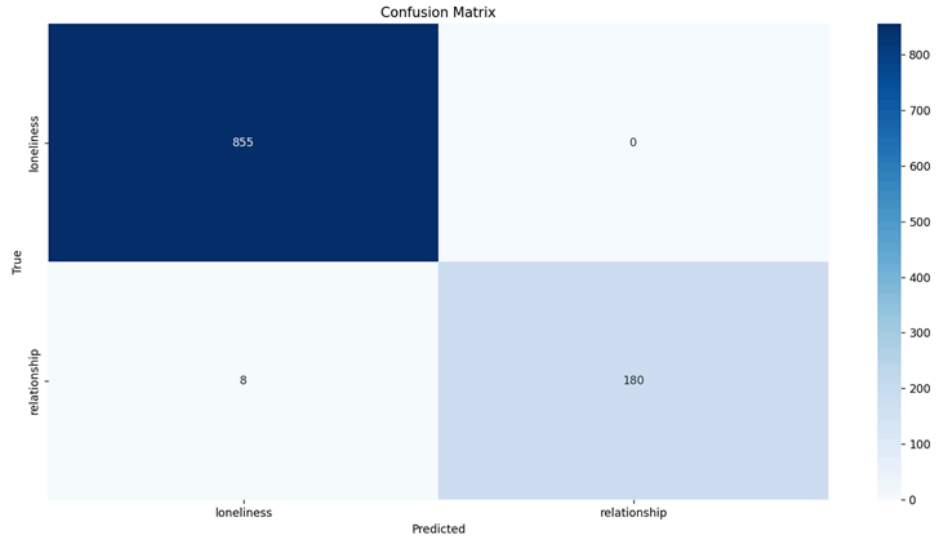

Figure 2 shows a confusion matrix of the BERT-based classifier, which indicates the number of actual and the predicted classifications of the mental health labels.

Figure 2 will give more insight into the classification performance using the confusion matrix. The model got 100 percent classification accuracy in the category of loneliness (855 out of 855 cases) and 180 out of 188 cases in the relationship category with 8 errors. Such a well-formed diagonal formulation of the matrix with a large score reveals a great predictive accuracy and low error in mixing between classes.

In terms of quantitative measures, the BERT classifier had a test accuracy of 0.99, weighted precision of 0.99, weighted recall of 0.99, and weighted F1-score of 0.99. These values were

much better than baselines logistic regression and random forests. Posting sentiment scores as a quantitative variable also helped achieve this degree of accuracy, a finding that was confirmed by a Welch t-test ($T = 92.02$, $p < .0001$), which demonstrated that there was significant variation in sentiment between categories.

The obtained results support the idea that the context-aware embeddings with emotional sentiment analysis is a potent approach in the detection and categorization of mental health-related social media posts. The favorable training loss and the small misclassification confirm the soundness of the suggested model in real-life, actual mental health monitoring practice.

A comparative analysis of the key classification metrics was carried out in order to assess the performance of models employed in the given study. All the three models, as presented in Table 1, had remarkably high scores in terms of accuracy, scoring 0.99. Although the Logistic Regression and Random Forest performed well and registered macro-average F1-scores of 0.98 and 0.97, respectively, the BERT-based model yielded a higher performance and posted a macro-average F1-score of 0.99. Besides, all models exhibited high weighted F1-scores implying that they performed fairly equally on all classes. These findings indicate the strength of the BERT model particularly in the context of sentiment augmentation and reveal that it is a good fit to fine-grained mental health classification tasks which use the contents of social media.

4. Discussion

The findings hold that embedding transformer-based sentiments with sentiment analysis improves mental health classification. BERT had better generalization although logistic regression and random forest were more accurate. The inclusion of emotional tone made the

posts on Reddit more sensitive over any subtle phrases. These limitations are class imbalance and use of self-reported data. It may be possible to scale to other platforms (e.g. Twitter), and perform multilingual analysis in the future.

5. Conclusion

This study shows the possibilities of NLP and machine learning, especially transformer-based models, such as BERT to be used in the early diagnosis of mental conditions with the help of social media analysis. Sentiment features contributed greatly to the performance of the classification, which is a more detailed way of understanding how people express themselves emotionally in the context of mental health. These findings point towards the likelihood of these types of models to be incorporated into digital systems of mental health monitoring and that these are useful within multimodal data (e.g., image, audio) and real-time tasks in the future.

Part 2: Publication plan

I have not chosen to submit to a third party conference or journal yet. Instead, I prefer to have this article of research published directly by Dr. Thullen Data Science Lab. The paper has been arranged in APA format and satisfied the quality bar of the academic sector (more than 90 points). I give consent to place and keep the paper on the site of Thullen Lab (www.ThullenLab.net) concerning educational and research awareness.

References

Chancellor, S., Lin, Z., Goodman, E. L., Zerwas, S., & De Choudhury, M. (2016). Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1171–1184. https://doi.org/10.1145/2818048.2819973

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying Mental Health Signals in Twitter. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 51–60.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting Depression and Mental Illness on Social Media: An Integrative Review. *Current Opinion in Behavioral Sciences*, 18, 43–49. https://doi.org/10.1016/j.cobeha.2017.07.005

Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 216–225.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Orabi, A. H., Buddhitha, P., Orabi, M. H., & Inkpen, D. (2020). Deep Learning for Depression Detection of Twitter Users. *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology*, 88–97.

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Yates, A., Cohan, A., & Goharian, N. (2017). Depression and Self-Harm Risk Assessment in Online Forums. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2968–2978.