Detecting Depression from Tweets with Neural Language Processing

Sijia Wen 1*

1Saint Edward’s School, Vero Beach, FL, 32968, USA
*Corresponding author’s e-mail:kathywl@steds.org

Abstract. As social media becomes a major part of everyday life, analyzing language on social media is becoming a potentially fruitful approach to discover depression patients. At a time when depression is starting to be taken seriously, sentiment analysis through social media should be studied further so that people can be treated in the early stages of depression. With the intention to detect depression behavior through studying language on social media, we built a classification model for depression detection by training on a Tweets dataset from Shen et al 2017. After optimizing hyper-parameters including learning rate and embedding dimension, the language classification model achieves the test accuracy of 98.94% and an F1 score of 99.04% which is higher than the best performance of 85% F1-measure achieved by Shen’s work. These results show that the method is effective and can be used in a wider range of unlabeled data to locate potential depressed users.

1. Introduction
Mental health is an important factor in maintaining social health and stability. It is closely related to the way people think, act, and feel. In the past several years, mental disorders have become one of the most common public health concerns in the world. Depression, a major contributor to global mental illness, influences the lives of millions of people of all ages. According to the World Health Organization, there are more than 264 million people in the world suffering from depression. Depression ruins families and causes an increase in the suicide rate.

Additionally, depressive disorders are the second worldwide main cause of years lived with disability (YLDs), a measurement of the burden of disease, and also the leading cause of disability-adjusted life years (DALYs). A measurement of the difference between current health status and an ideal situation where the entire population lives to the standard age and is in perfect health[3]. It was estimated that the number of people who claim to suffer from depressive disorders increased by nearly 18% from 2005 to 2015[3]. Traditionally, mental illness is diagnosed by face to face interviews with psychological doctors. However, it was estimated that 50% of people do not consult doctors when they have depression, which leads to further degradation of their situation[9].

Social media is a convenient tool for users to create and share content and has become an important part of people’s lives in recent years. Users can interact with each other and share ideas in a virtual community. Posts about people’s feelings, thoughts, and daily lives provide prodigious amounts of useful information about their mental states and social behaviors[1]. Twitter is one of the biggest social networking and micro-blogging platforms with an average of 330 million monthly active users. It contains posts from different countries, and its audience varies from various social stratus and interest groups. Since Twitter comprises many valuable opinions worldwide, it is the ideal platform to conduct our study. Besides, social media also transforms how people with depression behave. While
lethargy is some of the obvious symptoms, patients nowadays tend to share feelings more freely online. In this context, it is feasible to capture behavioral attributes and analyze the language used by applying machine learning and data mining techniques.

In this paper, a deep neural network architecture was applied to predict depression. The dataset used in this research was adopted from Shen et al, which includes tweets labeled as depressed and non-depressed ranging from 2009 to 2016. We chose PyTorch[8], an open-source machine learning library, to train a supervised learning algorithm for classification. PyTorch provides a Python package for high-level features like tensor computation with strong GPU and is commonly used in applications including computer vision and Natural language processing (NLP). Using PyTorch’s torch.nn module, a feed-forward neural network was built which is composed of an EmbeddingBag layer and a linear layer. The main contributions are:

1. Through optimizing the embedding dimension and learning rate, we attained an held-out test set accuracy of 99.45%, performing better than the 85% accuracy obtained by Shen et al[10].

2. Our well-trained model can be applied by mental health researchers and counselors to detect depression on large unstructured datasets in an automatic and unsupervised manner, enabling early detection and treatment of depression before symptoms worsen.

2. Related Works

[10] Shen et al. proposed a multi-model dictionary learning (MDL) method to detect depressed users on Twitter. They defined and extracted six depression, oriented feature groups, according to common offline and online behaviors that associated with depression. By comparing different classification methods, they found that MDL outperformed in three ways, including the highest F1 score of 85%. Shen et al. proposed a multimodal depressive dictionary learning method. Despite our work sharing the same Twitter dataset, our model differs in that we built a feed-forward neural network to train the model.

[2] Choudhury presented a way to detect depression in social media by analyzing language. They collected and classified their data which coming from questionnaires and self-reports. A Support Vector Machine(SVM) classifier was trained in their work and performed with 70% classification accuracy. While they used distinguished attributes to build an SVM classifier, this model was used in this research to learn through a well-labelled training dataset to identify depressed users.

[4] Guntuku et al. compared an array of previously published methods, which used to predict depression through social media. It recommended that additional studies are needed which integrate social media data collection with gold-standard structured clinical interviews. They mentioned that publicly accessible data is large enough to be sampled. However, users who are unaware of their diagnosis are unlikely to be captured. They also referred that the detection of mental illness through social media may achieve predictive performance between unassisted clinician evaluation and screening survey.

[7] Pak et al. built a sentiment classifier by analyzing the corpus that was formed by 30,000 tweets. They first presented a method for an automatic collection of a corpus that can be used to train a sentiment classifier. Then, they built a multinomial Naive Bayes classifier and trained two Bayes classifiers: one uses N-gram as a binary feature and one uses POS-tags to calculate posterior probability. They claimed that bigrams are useful to achieve the best performance and salience strategy which discriminates common n-grams provides better accuracy. However, once the dataset is large enough, their model may stop improving by simply increasing the size of the training data.

[5] Leis et al. used the Freeling Natural Language Processing tool to tag part-of-speech(POS) to analyze the usage patterns of grammatical categories. Helped by this technique, they successfully found the difference in language patterns between depression and normal users. Their work proves the possibility of depression detection on social media.
3. Depression Detection with Neural Networks

3.1. Data Collection and Processing
We adopt the data set from Shen et al[10], which consists of tweets from Tweeter’s API (2009 and 2016). There was a useful set of 11,877 tweets in the dataset, and 5,384 tweets were marked as depressed. Their dataset is useful because they classified it into depression datasets D1, non-depression dataset D2, and depression-candidate dataset D3. According to Shen et al., depression dataset D1 contains tweets between 2009 and 2016, and only the users who used the phrase: “(I’m/ I was/ I am/ I’ve been) diagnosed depression.” will be counted in this dataset. Non-depression dataset D2 consists of tweets in December 2016 from users who did not post the word “depress” in any of their tweets. Depression dataset D3 has much more noise. They used a looser pattern to select depressed users compared to D1. Users who post a small number of tweets containing the word “depress” will be counted in D3. Furthermore, they filtered and cleaned their data. Accounts that have more than 15,000 followers were removed from the dataset, since they might be the accounts managed by companies or organizations. Narrative contents were also removed because they do not represent the feelings and moods of users themselves. Tweet that mentions users’ past illnesses were also removed. Finally, they deleted the retweeted tweets which cannot present users themselves and removed the tweets with certain hashtags like #Factaboutme.

To process the datasets, we used json.load to deserialize the dataset, and extract the tweets for each user. A label was added before each tweet with number 0 or 1: 0 represents a non-depressed tweet, 1 stand for a depressed tweet. The resulting processed datasets were saved into a Python list. Since the dataset is sufficiently large, we used train_test_split function in Sklearn model selection to create two subsets: a training set with a size of 80% of all tweets and a test set with the remaining 20%. Once the dataset is prepared, a vocabulary is constructed from the training set, and a filter then removes words that are not in the vocabulary from the test set. The vocabulary provides an encoding from strings of words to vectors of (arbitrary) integers which can be further processed by PyTorch. After processing, we arrived at 5,384 training instances consisting of (1) a depressed and non-depressed Boolean label and (2) a long tensor representing the encoded tweet.

![Figure 1. Model Structure. The raw dataset was processed by grabbing the tweets of each user and labelling them by 1 or 0 (depressed/non-depressed). The data then passed through the embedding layer where a look-up table is created in which each row represents an embedding of a word. After this, the linear layer applied a linear transformation to the incoming data and yielded the output for the model.](image-url)
3.2. Model Architecture
Most of the learning tasks were concentrated in the traditional single-layer net-work. However, a multi-layer network is also suitable to apply in the real world example which may contain potential exceptions and noises. In this paper, a feed-forward neural network was used to build our model. Figure 1 shows the structure of our neural network. The term “feedforward” suggests that information flow through the input nodes, then through the hidden layers, and finally to the output layer without any feedback connections. In our study, the model architecture mainly consists of an embedding layer followed by a linear layer. We sent words in sentences to a word lookup table which translates words into values and saves into an array. After that, the embedding layer is created by nn.EmbeddingBag function in PyTorch library. In this layer, all words are treated as independent entities. By embedding the tweets, semantic information of a word will be preserved [6]. Next is the linear layer. This layer applies a linear transformation to the incoming data:

$$y = xA^T + b$$  \hspace{1cm} (1)

In nn.Linear function, the dimension of x is the embedding dimension, and y is a 0/1 scalar representing whether this instance is depressed/non-depressed. After the structure of the model is defined, we used cross-entropy loss to calculate the performance of our classification model. Cross entropy loss measures the distance between the predicted value and the actual label.

$$J(w) = \frac{1}{N} \sum_{n=1}^{N} H(y_n, \hat{y}_n) = -\frac{1}{N} \sum_{n=1}^{N} [y_n \log \hat{y} + (1 - y_n) \log (1 - \hat{y})]$$ \hspace{1cm} (2)

where $\hat{y}_n$ is:

$$\hat{y}_n = \frac{1}{1+e^{-wx}}$$ \hspace{1cm} (3)

Here we defined J as the cross-entropy loss between the empirical distribution ($y_n$) and the softmax distribution defined by $\hat{y}_n$. In our case, $y_n$ is an indicator variable of depressed or non-depressed. If n is depressed, $y_n$ is 1. If n is not depressed, $y_n$ is 0. $\hat{y}_n$ is the resulting softmax probability which is the output of the neural network model, and $x_n$ is the result of the embedding bag which is a matrix of the vector of words. We optimized the model using Stochastic Gradient Descent (SGD). In a gradient descent framework, the slope is descending to reach the lowest point on a surface. However, it is slow on massive datasets. To avoid this problem, we used Stochastic Gradient Descent(SGD) where SGD only iterates one random example each time from the entire data set. By converging faster, SGD is able to perform updates more frequently when calculating over big data. The learning rate is a configurable hyper-parameter which determines how far along the current gradient each iteration should step. Adjusting both the learning rate and the embedding dimension enables the model to reach the highest accuracy. Lastly, we completed our computation by calling the backwards function from PyTorch’s automatic differentiation package in order to compute derivatives automatically.

3.3. Results
We first set the number of epoch to 80, performing a grid search over four different values of the embedding dimension and four different values of the learning rate.
Figure 2. Test and training loss. The results of the loss using different values of parameters are shown above. By observing the test loss plot, we excluded the models with a learning rate of 1 (the red lines) and 0.1 (the green lines) since they all presented an overfitting trend.

Figure 3. Plot of the best model with an embedding dimension of 64 and a learning rate of 0.01 in 800 epochs range.

Since dealing with an imbalanced dataset, the F1-score was measured for the test dataset:

\[
F1 = 2 \times \frac{(precision \times recall)}{(precision + recall)}
\]

F1-score calculates the harmonic mean of precision and recall. In the code, the f1_score method was used in the Sklearn library. Parameters y_true (the correct target value) and y_pred (the predicted targets) were set to the label and the output accuracy of the test dataset.

The result of the grid search over the embedding dimension and learning rate is in Figure 2. After comparing, we discovered the best pair to train the model: embedding dimension of 128 and learning rate of 0.01. In this case, the dashed red line showed a decrease in loss without any overfitting. As a result, the model achieves a high accuracy of 98.94% and an F1 measure of 99.04% on the test dataset.

In order to check if the model was stopped at the right point and help mitigate overfitting, we retrained the best model over a range of 800 epochs. The result is plotted in Figure 3. In this graph, we observed that the test loss of the model did not show an obvious further improvement after 80 epochs. As a result, we early stopped our final model at 80 epochs.

4. Conclusion

In this paper, the problem of depression detection was treated by using computational linguistic methods. A neural network language model was established to attain a high accuracy of 99.45% and F1-measure of 99.4% through training. It outperforms prior state-of-the-art methods that attains an F1-score of 85% [10]. Our work is important because it can be deployed for automatically early-stage detection of depression by identifying potentially depressed users through text recognition on social
media. It also provides a perspective of using PyTorch text classification on detecting depression for future research. By discovering depression through analyzing online behaviors, our model produces better patient outcomes, reduces healthcare costs, and enables people to discover potential psychological problems and treat their illness in the early stage.

References
[1] J. Clement. Twitter: monthly active users worldwide, Aug 2019.
[2] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. Icwsm, 13:1–10, 2013.
[3] Alize J Ferrari, Fiona J Charlson, Rosana E Norman, Scott B Patten, Greg Freedman, Christopher JL Murray, Theo Vos, and Harvey A Whiteford. Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. PLoS med, 10(11):e1001547, 2013.
[4] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. Detecting depression and mental illness on social media: an integrative review. Current Opinion in Behavioral Sciences, 18:43–49, 2017.
[5] Angela Leis, Francesco Ronzano, Miguel A Mayer, Laura I Furlong, and Ferran Sanz. Detecting signs of depression in tweets in spanish: behavioral and linguistic analysis. Journal of medical Internet research, 21(6):e14199, 2019.
[6] Toma`s Mikolov, Stefan Kombrink, Luk`aˇs Burget, JanˇCernockˇy, and SanjeevˇKhudanpur. Extensions of recurrent neural network language model. In 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 5528–5531. IEEE, 2011.
[7] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In LREc, volume 10, pages 1320–1326, 2010.
[8] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary De Vito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlche-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32, pages 8024–8035. Curran Associates, Inc., 2019.
[9] Claire Rondet, Isabelle Parizot, Jean Sebastien Cadwallader, Jacques Lebas, and Pierre Chauvin. Why underserved patients do not consult their general practitioner for depression: results of a qualitative and a quantitative survey at a free outpatient clinic in paris, france. BMC family practice, 16(1):1–13, 2015.
[10] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. Depression detection via harvesting social media: A multimodal dictionary learning solution. In IJCAI, pages 3838–3844, 2017.