Depression Detector in Tweets using TF-IDF and BOW

Muhammad Athallah Rizki Putra
Telecommunication Engineering
Bandung Institute of Technology
Bandung, Indonesia
18117002@std.stei.itb.ac.id

Muhammad Hanif Naufal Eka Wiratama
Telecommunication Engineering
Bandung Institute of Technology
Bandung, Indonesia
18117027@std.stei.itb.ac.id

Hafizh Mulya Harjono
Telecommunication Engineering
Bandung Institute of Technology
Bandung, Indonesia
18117030@std.stei.itb.ac.id

***Abstract*—Depression is a leading cause of mental ill health, which has been found to increase risk of early death. Moreover it is a major cause of suicidal ideation and leads to significant impairment in daily life. For some people, sometimes they express their feelings in social media. The advent of social media has resulted in significant user data being available for sentiment analysis of text. This paper aims to apply machine learning models on Twitter tweets for conducting depression. Individual tweets are classified as depressive or non-depressive, based on a curated word-list to detect depression tendencies. In the process of class prediction, term frequency-inverse document frequency (TF-IDF) and bag of words (BOW) have been used. The results have been presented using the primary classification metrics including F1-score, accuracy, recall, and precision.**

Keywords—depression, tweets, sentiment analysis, TF-IDF, BOW

I. Introduction

Depression is a common mental disorder and one of the main causes of disability worldwide. Globally, there are about 450 million people who are affected by depression. Mental disorders including depression comprise a broad range of problems, with different symptoms. However, they are generally characterized by some combination of abnormal thoughts, emotions, behavior, and relationships with others [1]. South Asian countries have the highest prevalence of common mental disorders (CMDs) globally, with nearly 28,4% [2].

However, naturally, before an individual with depression receives treatment, this disorder must be detected. Many patients do not receive an earlier depression diagnosis in consultation with general practitioners, with roughly 50% of the cases detected [3]. Thus, it is important to find other ways to detect depression apart from clinical diagnosis with psychological and psychiatry doctors.

It is observed that some people like to express their emotions in social media by sending texts or pictures. Twitter is one social media where people from anywhere in the world share their aspects of life and opinions on current hot topics. As Twitter has become a very popular platform for communication or to tell private stories, it will be very potential to detect depression from its content [4][5].

In this research, a machine learning approach is used to detect depression by analyzing the social media posts of users. Twitter posts have been considered to convey the model. There are lots of parameters to be acknowledged to indicate depression of a user. Most of the users express their emotional state through tweets. To analyze the collective data from Twitter, effective machine learning classification techniques are used here.

Emotion AI has been applied to the collected and preprocessed Tweet data, which will classify whether tweets contain a depression statement or not. Supervised learning is the machine learning task which involves providing the algorithm with labeled datasets, which is then used to learn model parameters. This paper implements TF-IDF and BOW classifier for detecting Tweets which show signs of depression.

II. Methodology

The workflow starts with a data collection step, which utilizes a Python library called GetOldTweets3 for the generation of datasets. Authors also collect tweets from the available datasets on the internet. Following the collection of datasets, the data is preprocessed through tokenization, stemming, and stop word removal. After this, the text classifier is trained on the processed text data from Twitter, in the training phase. In the testing phase, the class prediction is made on the test dataset to identify potential Tweets demonstrating depression tendencies. Last, using the model that has been trained and tested, the prediction is done by entering some tweets or texts to see the result whether that tweets or texts contain depressive meaning.

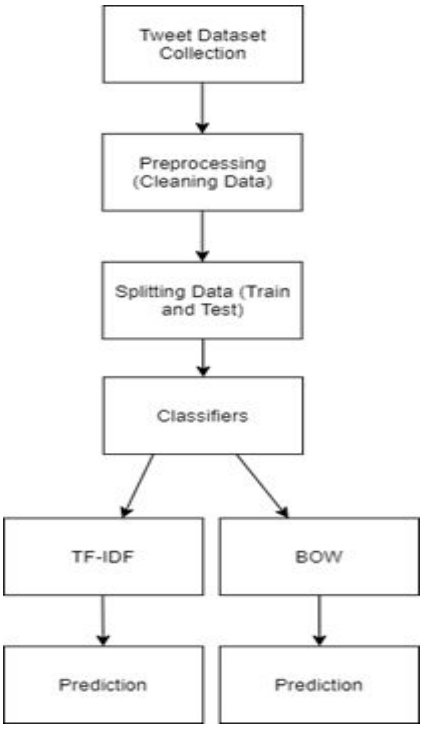

Fig. 1 Workflow Chart

A. Dataset Collection

Some of the labeled tweets dataset are collected from https://www.kaggle.com/kazanova/sentiment140. Authors took 8000 depressive labeled tweets. The other tweets are obtained by using GetOldTweets3 to collect depressive tweets using "depression" and "depressed" keywords, 2314 tweets were collected. For the tweets that were scrapped by using GetOldTweets3, the authors need to manually label the tweets. The authors put "0" on the label column to indicate non-depressive tweets, and put "1" on the label column to indicate depressive tweets. After that, the authors combined all the datasets to a new CSV (comma-separated value) file containing messages (tweets) and labels.

	text	label
0	just had a real good moment. i miss...	0
1	is reading manga http://plurk.com/p/mzp1e	0
2	@comeagainjen http://twitpic.com/2y2lx - http:...	0
3	@lapcat Need to send 'em to my accountant tomo...	0
4	ADD ME ON MYSPACE!!! myspace.com/LookThunder	0

Fig. 2 Example of Collected Dataset

B. Preprocessing

The CSV file is read and several data preprocessing steps are performed on it. Natural language processing algorithm below, sequentially, has been utilized for preprocessing methods applied on the extracted data:

- Tokenization: Tokenization is a process of dividing a string into several meaningful substring, such as units of words, sentences, or themes. [6] In this case, the messages (tweets) column of the csv file is extracted and is converted into individual tokens, which are an array of words. This step has to be done because natural language processing (NLP) can only read each word and give them value in token form.
- Removing Stop Words: NLTK library has a set of stop-words which can be used as a reference to remove stop-words from the tweet. The stop words contain non-relevant meaning to the context, hence it should be removed from the tokenized data.
- Stemming: Stemming involves removing affixes in each word, if there are no affixes, then the word will not change. A more accurate definition is that stemming will return a word to a standard form, which may differ from its dictionary form. This would help us to group similar words together, such as "plays", "playing", and "played" will be stemmed as "play"; "is", "are", and "am" will be stemmed as 'be'. Porter stemmer algorithm is used for this process.

	text	label
0	real good moment miss much	0
1	read manga e	0
2	lx	0
3	need send em accountant tomorrow oddly even re...	0
4	add myspace myspace com lookthunder	0

Fig. 3 Example of Dataset After Preprocessing

After all these pre-processing steps, TF-IDF and Bag of Words classifier are formed. TF-IDF and Bag of Words give value to each word by calculating the number of occurrences of each word, which is then used as a feature to train the data.

C. Data Splitting

Training the model requires two separate sets: the training set and the test set. The authors only had a few collected datasets, so almost all data are allocated as the training set to prevent overfitting and low accuracy. Authors used 98% of the data as the training set and the rest as the test set.

D. Classifier

Authors used two methods to classify data: term frequency-inverse document frequency (TF-IDF) as the default method and bag of words (BOW). The TF-IDF weight is often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word, or more generally a token such as 2-gram, is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the TF-IDF weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. One of the simplest ranking functions is computed by summing the TF-IDF for each query term; many more sophisticated ranking functions are variants of this simple model. TF-IDF can be successfully used for stop-words filtering in various subject fields including text summarization and classification [7].

The BOW model is a simple representation used in NLP and information retrieval (IR), also known as a vector space model. In this model, a text in the form of a sentence or document is represented as a multiset bag of words contained in it, regardless of word order and grammar but still maintains its diversity. Another definition for BoW is a model that learns a vocabulary from the entire document, then models each document by counting the number of occurrences of each word [8].

E. Prediction

The last step is doing prediction using the classifiers that have been created. Authors put some tweets or sentences in the predictor of two classifiers. Depressive tweets and non-depressive tweets were tested to see the prediction results.

III. Results and Discussion

In this section, we present the experimental results obtained from the machine learning models, as explained before. We start by collecting tweets data, followed by labeling the data as depressive and non-depressive tweets. Next, we train the TF-IDF model and see the accuracy. Then, if the accuracy is good enough, we test the model by entering some tweets to see whether the model has correct inference of the tweets.

```
{'depression anxiety': 135,
 'anxiety depression': 132,
 'depression https': 84,
 'mental health': 51,
 "depression n't": 43,
 'face emoji': 43,
 'crying face': 41,
 'emoji face': 39,
 'depression emoji': 37,
 'depression http': 37,
 'emoji loudly': 33,
 'loudly crying': 33,
 'mom depression': 33,
 'emoji heavy': 31,
 'smiling face': 31,
 'great depression': 31,
 'mental illness': 31,
```

Fig 4. 2-grams with depressive label, sorted from the highest TF-IDF weight

By default, after tokenization the 2-gram datas are weighted by TF-IDF, but the BOW model is also supported. The depressive labels and non-depressive labels are weighted separately. Fig. 2 shows some 2-gram with depressive labels. We could say that sentences that have phrases shown above have a high chance to be classified as depressive. The TF-IDF weights and model then are used to classify tweets.

```
pm = process_message('I\'m depressed')
sc_tf_idf.classify(pm)
```

```
True
```

```
pm = process_message('Depression are the worst')
sc_tf_idf.classify(pm)
```

```
True
```

```
pm = process_message('Lately I have been feeling unsure of myself as a person & an artist')
sc_tf_idf.classify(pm)
```

```
False
```

Fig 5. Classification results

To classify a tweet, at first it takes one string as an input for inference which goes through all the steps of preprocessing such as removing tokenization, stemming, and stop words removal. After that, feature extraction tweets are analyzed using the TF-IDF mathematical model. Sentiments are determined by calculating the score of each word and the probability of each word occurring in the dataset. Tweets are validated according to positive and depressive scores of sentences with the aim of determining the accuracy of these analyzers. Bag of Words classifiers are also implemented to give a comparison of their accuracies. The examples are shown in Fig 2.: depression is detected in "I'm depressed", and so on. To highlight the examples, the first two examples shown in Fig 5. have the word 'depressed' and 'depression' which are seemingly present in the list of 2-grams with depressive labels, hence the depressive label.

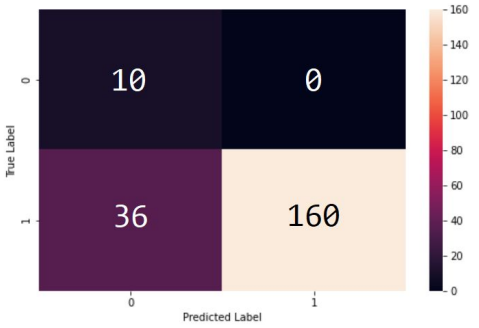

Fig 6. Confusion matrix

Fig. 6 shows the confusion matrix from test data. It consists of two rows and two columns that shows true positive, false positive, false negative, and true negative parameters. The vertical axis shows the true labels from test data, and the horizontal axis shows the predicted labels. True positive is defined as the number of depressive tweets that are successfully classified as depressive, and true negative is defined as the number of non-depressive tweets that are successfully classified as non-depressive. On the other hand, false positive is defined as the number of non-depressive tweets that falsely classified as depressive, and false negative is defined as the number of depressive tweets that are falsely classified as non-depressive. From Fig. 6, it is shown that the TF-IDF model has a true positive value of 10, a true negative value of 160, false positive value of 0, and false negative value of 36. The true negative value is much higher than the true positive value due to the different size between depressive and non-depressive labeled tweets in the dataset. The model's false negative value indicates that the model still can not detect depression from some tweets.

Table I. Performance Parameters

Proposed Classifiers	TF-IDF	BOW
Accuracy	86,5%	83,2%
Precision	92%	100%
Recall	46%	28%
F-Score	61,3%	43,7%

The accuracy of the TF-IDF and BOW classifiers is different. BOW analyzes the words by calculating the occurrence probability of each word, while TF-IDF analyzes by multiplying how many times a token (in this case is a 2-gram) appears in the dataset, and the inverse frequency of the token across the dataset. TF-IDF is using logarithmic normalization that makes the classifiers get better results. The accuracy parameter is defined as the fraction of correctly classified instances among all data, in this case is formulated by the sum of true positive and true negative, divided by the number of test data. The precision parameter is defined as the fraction of relevant instances among the retrieved instances, in this case is formulated by true positive value divided by the number of depressive-labelled datas. The recall parameter is defined as the fraction of the

total amount of relevant instances that were actually retrieved, in this case is formulated as the true positive value divided by the number of data that are classified as depressive-labelled. The F-score is calculated from precision and recall value, formulated by

$$F_{score} = 2\frac{precision \cdot recall}{precision + recall}. \quad (2)$$

The highest value of F-score is 1, indicating perfect precision and recall; and the lowest value of F-score is 0, indicating zero precision and zero recall. Table I shows the performance parameters for the models. The TF-IDF model has a higher value of accuracy and F-score than the BOW model, so we could say that the TF-IDF model performs better.

The score of each parameter and the inference results for both models are not optimum and do not show their maximum potentials. The dataset used for training do not have the same amount of depressive and non-depressive dataset, also the tweets contain many slang words and ignore grammar that make the preprocessing challenging. This calls for further research on this area to improve the dataset quality by scraping more tweets with better data format or to improve the preprocessing algorithm. The authors also suggest using a dataset with a larger amount of data to represent the real world better.

IV. Conclusion

This paper proposed a model that takes sentences/tweets and analyzes them whether they contain depressive meaning or not. The collected tweets are analyzed and classified by the model as depressive or non-depressive. For the result, we evaluated the accuracy of our model which was 86,5% for TF-IDF classifier and 83,2% for BOW classifier. The confusion matrix is also shown in Fig. 6. The model proposed could still be improved by using a better dataset or by using an improved preprocessing algorithm to achieve a better accuracy. Moreover, every social networking site can implement this model on their respective platforms which will help to detect the depressed individual even more.

References

[1] "Mental Disorders," [Online]. Available: https://www.who.int.

[2] S. Naveed, A. Waqas, A. M. D. Chaudhary, S. Kumar, N. Abbas, R. Amin, N. Jamil and S. Saleem, "Prevalence of Common Mental Disorders in South Asia: A Systematic Review and Meta-Regression Analysis".

[3] Mann, Paulo, A. Paes and E. H. Matsushima, "See and Read: Detecting Depression Symptoms in Higher Education Students," Arxiv, vol. I, no. 2, p. 1, 2020.

[4] Rajput, A. E. and S. M. Ahmed, "Making a Case for Social Media Corpus for Detecting Depression," in the International Academic Conference on Humanities and Social, Leipzig, 2018.

[5] P. Aurora and P. Arora, "Mining Twitter Data for Depression Detection," IEEE, vol. III, no. 7, p. 187, 2019.

[6] M. Deshpande and V. Rao, "Depression Detection using Emotion Artificial".

[7] "What does TF-IDF mean?," [Online]. Available: http://www.tfidf.com/#:~:text=What%20does%20tf%2Didf%20mean,in%20a%20collection%20or%20corpus.

[8] W. T. H. Putri and R. Hendrowati, "Penggalian Teks dengan Model Bag of Words terhadap Data Twitter".