2020 Information Technology International Seminar (ITIS)
Surabaya, Indonesia, October 14-16, 2020

Twitter Sentiment Analysis of Covid-19 Using Term Weighting TF-IDF And Logistic Regresion

Imamah
Faculty of Engineering
University of Trunojoyo Madura
Bangkalan, Indonesia
i2m@trunojoyo.ac.id

Fika Hastarita Rachman
Faculty of Engineering
University of Trunojoyo Madura
Bangkalan, Indonesia
fika.rachman@trunojoyo.ac.id

2020 6th Information Technology International Seminar (ITIS) | 978-1-7281-7726-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ITIS50118.2020.9320958

***Abstract*—Covid-19 attack world population, and has brought much impact in all aspects of life. Stay at home and doing less in terms of social interactions. This can have a negative effect on mental health, so in this study we use sentiment analysis to know about mental health through public opinion on Twitter. Dataset which used in this study in this study is covid-19 tweets collected at 30 April 2020. Essentially, this dataset consists of 355384 tweets reviews. Covid-19 tweets will classify with the Logistic Regression method. Based on this research, the accuracy of the covid-19 tweeets sentiment classification is 94.71%.**

Keywords—coronavirus, twitter, sentiment analysis, prepocessing, positive, negative.

I. INTRODUCTION

A sentiment analysis system usually involves of opinion mining and sentiment analysis. The prior purposes to classify the aspect and opinion revealed in reviews of data which is obtained by crawling techniques on social media or website [1],[2], and the last one to gather polarity score to classify sentiment [3],[4]. In the other words, Sentiment analysis is used to examine feelings, views, emotions, expression, beliefs, attitude, and opinion [5].

Nowadays people using social media like Twitter to express their opinion. According to *Internet Live Statistics* meanwhile 2013 the amount of tweets sent every day has stretched 500 million[6]. Some studies have been issued on the aptitude of twitter to forecast the whole thing from the stock market[7] to elections[8]. In this study, we collect data reviews from Twitter to know about public opinion in worldwide about Covid-19. As we know that Covid-19 attack World population, and has brought much impact in all aspects of life. We should stay at home and doing less in terms of social interactions. This can have a negative effect on mental health of people in the word, so in this study we use sentiment analysis to know about mental health through public opinion on Twitter.

With the sentiment analysis of covid 19 data from Twitter, we can see the mental health of people in the world, whether they still feel safe or are they in the excessive worry stage. Safe conditions can be observed from the neutral and positive sentiment values, while the worry condition can be seen from the negative sentiment values.

II. RELATED WORK

This section will provide a literature study about sentiment analysis. Cahyo et all shown an examination of the sentiment analysis of mobile banking opinions using TF-IDF and cosine Similarity[9]. In 2019, Soumya using Malayalam tweets dataset with machine learning techniques to analyze the sentiment. The tweets are classified into positive and negativeusing Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest (RF) [10]. Abyan collect data reviews of the body shop tea tree oil on Twitter with positive and negative sentiments and with the accuracy level of Tea Tree Oil sentiment analysis models on female daily at 61.51%[11]. The other research performed the sentiment analysis of Go- Pay tweets. The researcher using lexicon to classify the sentiment into positive and negative label. The accuracy of linear kernel SVM is 89.17% and the polynomial kernel is 84.38% [12].

TF-IDF does not contain the number of occurrences of words alone as in BOW. But TF-IDF also looked at the important and less important words of the document. Previous research [13] compared selection features using BOW with TF-IDF. The result is that TF-IDF is better than BOW, so in this study using the TF-IDF concept in its selection features.

The choice of classification method is very important to obtain proper accuracy and in accordance with existing data. According to research [14], if the test involves positive and negative predictive values, the appropriate algorithm is binary logistic regression. In this study the predictive data are negative, positive and neutral, so we use the logistic regression method.

III. PROPOSED METHODS

Term Weighting TF-IDF and Logistic Regression is a proposed method to classify neutral, positive and negative sentiment. The description of our proposed work as seen in the Fig. 1.

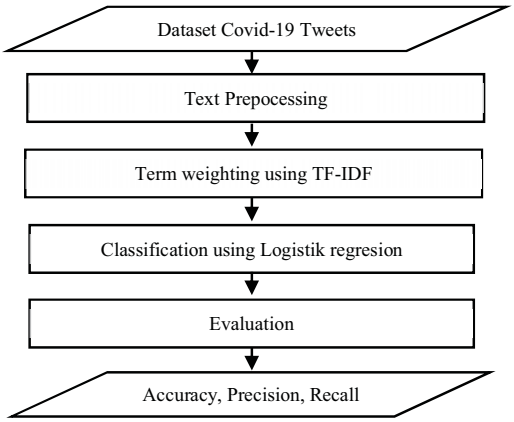

Fig. 1. Proposed methods of sentiment analysis.

In accordance with Figure 1, the initial stages are doing text prepocessing on the crawled data from Twitter. Preprocessing results in the form of terms. The term is then selected for features using TF-IDF to produce vectors. Data from the TF-IDF method were classified using the logistic regression method. Then the evaluation process is carried out with recall, precision and accuracy.

A. Dataset

Dataset which used in this study in this study is covid-19 tweets collected at 30 April 2020. Essentially, this dataset consists of 355384 tweets reviews. Example of dataset as seen at Fig. 2.

	created_at	text
355379	2020-04-30T23:59:58Z	dear nycmayorthe city millions spend interfait...
355380	2020-04-30T23:59:59Z	middle covid sales surge march kogan delivery ...
355381	2020-04-30T23:59:59Z	lockdown covid continues improved air quality ...
355382	2020-04-30T23:59:59Z	best answer covid wemournthemall httpstcoqrrvahgb
355383	2020-04-30T23:59:59Z	wfxtmalini cmichaelgibson boston three positiv...

Fig. 2. Covid-19 Tweets Dataset

B. Text Prepocessing

Text preprocessing is part of text mining to find patterns and to get valuable information for a specific purpose[15]. Based on the anomalies structure or unstructured of data text, the text processing is required some to develop data text becomes structured[16]. The steps of text preprocessing are tokenizing, filtering, and stemming.

C. Term Weighting TF-IDF

Term weighting is the technique that gives weights (values) to separately term in a document[17]. Three aspects influence the scheme of term weighting is term frequency(TF), inverse document frequency(IDF), and normalization.

In this study, TF- IDF is used as the feature weighting method. TF- IDF determine a weight of the term with two factors:

- Calculating the number of term *j* occurrences in document *i* or also known as term frequency denoted by $tf_{i,j}$.
- Calculating the frequent it arises in the entire document collection or also known as document frequency denoted by $df_{i,j}$.

Term weighting TF-IDF is considered formed on the following formula:

$$Tf.IDF = TF_{i,j} \; x \; IDF_{i,j} \; = TF_{i,j} \; x \; log\frac{N}{DFj} \qquad (1)$$

where,
N = number of documents in collections
TF = term frequency
IDF = inverse document frequency.

D. Classification using Logistic Regresion

Sentiment analysis is a process of classifying text into positive, negative and neutral. One of the method which is widely used to do supervised sentiment analysis is Logistic regression. The classification process using this method is done by extracts real-valued features from the input, multiplies each by a weight, sums them, and passes the sum through a sigmoid function to generate a probability. A threshold value is used to make a decision.

Logistic regression able to classify sentiment into two class with positive and negative label or multiple classes using multinomial logistic regression. In this study, Logistic regression classifier is used to train and test dataset which has three class (positive, neutral and negative) as labels. A weighting scheme is adapted for separately class, which is used to adjust the probability since the training data is unbalanced. [18].

E. Accuracy Testing

Accuracy testing is conducted to know the performance of proposed model in classifying the sentiment into three class labels (positive, negative and netra). The measurement techniques used for evaluation the proposed method is confusion matrix, with three classes. Confusion matrix will arrange actual class results and predictive classes. The final results of this research were constructed on the precision, recall, and f-measure which are usually used in evaluation metrics in sentiment classification research. Precision measures the perfection of a classifier, and recall dealings the wholeness or compassion of a classifier. The mixture of precision and recall is measured by the f-measure, which is the weighted harmonic mean of precision and recall.

IV. Result and Discussion

In this section, we the evaluation results of proposed method of sentiment classification on covid19 tweets dataset.

A. Prepocessing Data

Covid19 tweets dataset is unstructured and contain of noises. The dataset need to preprocessing in several stages are cleansing, eliminating numbers, emoticons, punctuation marks, case folding, filtering and tokenizing. The tools used I this research is Sklearn, Pandas and many more library of python for data preprocessing which runs on Jupyter Notebook. Example of preprocessing are shown in TABLE I.

TABLE I. Prepocessing Data

Original Tweets	***Prepocessing output***
Microsoft sees digital reboot from pandemic, profits up #COVID19 https://t.co/ruU6qomnfF	microsoft sees digital reboot from pandemic profits up covid httpstcoruuqomnff

B. Term Weighting

The words which is a result of preprocessing data will be converted into numeric using term weighting. This process aims to calculate the weight of each word that will be used as a feature, the more documents will be processed then the more features. At this stage there are two part of process namely TF (Term Frequency) and IDF (Inverse Document Frequency).

TF is the number of words that appear in each document. IDF is the number of document values in each word that are inversely proportional, it means if a word rarely appears in a document, the IDF value is greater than words that often appear. The result of weighting the word with TF-IDF is the multiplication of the values of TF and IDF which will produce less weight if the word often appears in every document in the

collection, on the contrary the weight of TF-IDF will be greater if the word rarely appears in each document in collection. By using TF IDF, you can also analyze the most tweeted words as in Fig. 3 and Fig. 4.

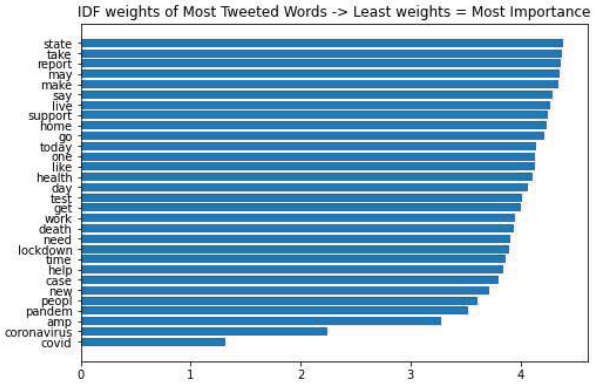

Fig. 3. IDF weights of most tweeted words

It can be seen in the graph in Fig. 3 that the terms that dominate the document in the top 10 terms are terms with neutral and positive sentiment values, for terms that have negative sentiment are below. But on the TF graph (Figure 4), it can be seen that the terms that have a "negative" sentiment like "death" are higher in value than those with a positive sentiment like "health".

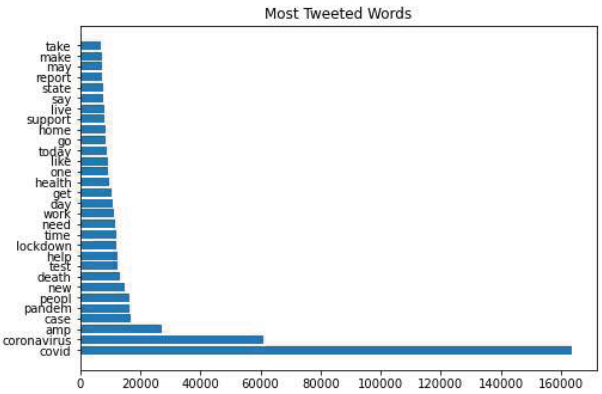

Fig. 4. Most Tweeted words using Term Frequency

C. Data Labelling

The next steps is data labelling which purposes to give positive, neutral and negative labels on Covid-19 Tweets dataset. In this study, we used polarity score to compute score of sentiment. Covid-19 reviews will be positive labelled if a score larger to 0. Otherwise if a score is less than to 0, it will be negative labelled. Covid-19 reviews will be neutral labelled when the score is equal to 0. The source code of sentiment polarity is shown in Fig. 5.

```
polarity_score = []

for text in df['text']:
    blob = TextBlob(text)
    for sentence in blob.sentences:
        if sentence.sentiment.polarity > 0.0:
            polarity_score.append('positive')
        elif sentence.sentiment.polarity < 0.0:
            polarity_score.append('negative')
        else:
            polarity_score.append('neutral')
```

Fig. 5. Source code to find polarity score fo sentiment

Total covid-19 tweets dataset is 355384 which includes 213122 neutral sentiment, 100683 positive sentiments and 41579 negative sentiments. Fig. 6 summarizes the statistics of this dataset.

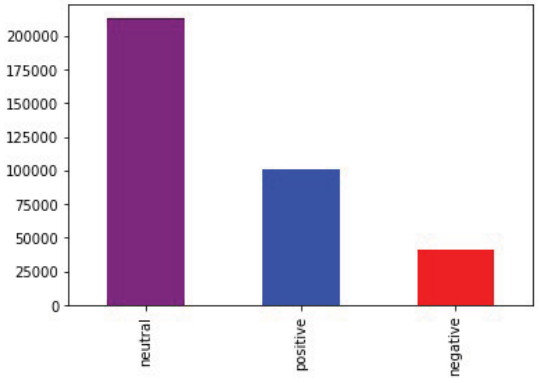

Fig. 6. Sentiment Classification of Covid-19 Tweets Dataset.

D. Data Visualization using Wordcloud

Wordcloud is used to display the words that often appear in the document. Here is the result of wordcloud visualization for covid-19 tweets. Fig. 7. describes the words that most appear which has positive label. The Examples of the most common words health, good, need, great, help.

Fig. 7. Wordcloud of positive sentiment.

Meanwhile Fig. 8 shows the words that most appear which has negative label, for example are cases, lockdown, bad, pandemic.

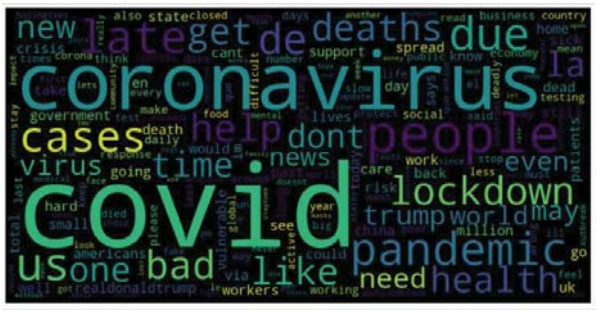

Fig. 8. Wordcloud of negative sentiment.

E. *Sentiment Classification Model*

The step which is so important before data modelling, we need to look the variable target as seen at Fig. 9 .

	created_at	text	target
355379	2020-04-30T23:59:58Z	dear nycmayorthe city millions spend interfait...	0
355380	2020-04-30T23:59:59Z	middle covid sales surge march kogan delivery ...	0
355381	2020-04-30T23:59:59Z	lockdown covid continues improved air quality ...	1
355382	2020-04-30T23:59:59Z	best answer covid wemournthemall httpstcoqrrvahgb	2
355383	2020-04-30T23:59:59Z	wfxtmalini cmichaelgibson boston three positiv...	2

Fig. 9. Variable target on Dataset.

After process of data modelling, the next step is splitting data into data train and data test. Data train will use for modelling, and data test will use for a model tester to evaluate unseen data. In addition it can be used to measure the capabilities of the proposed model.

Model evaluation is needed to know the performance of the selected method to classify covid-19 tweets. Confusion matrix with a size of 3x3 is used to measure accuracy of the proposed method as seen at TABLE II.

TABLE II. CONFUSION MATRIX OF COVID-19 TWEETS DATASET

Actual Class	Predict Class		
	Positive(P)	Negative(N)	Neutral(Ne)
Positive(P)	8104	1475	848
Negative(N)	82	53086	165
Neutral(Ne)	515	1611	22960

```
accuracy 0.9471444972199086
Precision 0.9471444972199086
Recall 0.9471444972199086
F1 Score 0.9471444972199086
              precision    recall  f1-score   support

           0       0.93      0.78      0.85     10427
           1       0.95      1.00      0.97     53333
           2       0.96      0.92      0.94     25086

    accuracy                           0.95     88846
   macro avg       0.94      0.90      0.92     88846
weighted avg       0.95      0.95      0.95     88846

Convolution matrix
[[ 8104  1475   848]
 [   82 53086   165]
 [  515  1611 22960]]
```

Fig. 10. Accuracy measure of Covid-19 Dataset.

By using Sklearn Python library, we achieved accuracy, precision, and recall as seen in the Figure 10.

Based on the Fig. 10, the evaluation model using logistic regression is 94,71%. With a percentage of accuracy 94.71%, it can be seen that the Logistic Regression method can be used well in sentiment analysis for COVID-19 data.

V. CONCLUSION

In this study, an experiment involving a term weighting and logistic regression to sentiment classification of covid-19 tweets. The use of TF-IDF in the extraction process had a significant impact on the classification performance of the sentiment analysis. The accuracy of system is 94,71%. Most tweet about covid-19 could be shown, so we know that most people giving the neutral sentiment, and the number of negative sentiment is lowest. So that it can be concluded that the mental health of people in the world is still neutral and positive. Even though there are people who are in the worry stage because they give excessive negative sentiments.

Future work could be done by improved the selection feature using n-gram, bi-gram, tri-grams to obtain better classification performance, rendering it as an effective technique in the extraction process of online documents or contents.

ACKNOWLEDGMENT

This research has funded by Faculty of Engineering, University of Trunojoyo Madura. The author wants to thank you for the opportunity that has been given.

REFERENCES

[1] P. Liu, S. Joty, and H. Meng, “Fine-grained opinion mining with recurrent neural networks and word embeddings,” in Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1433–1443.

[2] W. Wang, S. J. Pan, D. Dahlmeier, and X. Xiao, “Coupled multi-layer attentions for co-extraction of aspect and opinion terms,” in Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017, pp. 3316–3322.

[3] Y. Wang, M. Huang, L. Zhao, and E. Al, “Attention-based lstm for aspect-level sentiment classification Methods,” in Proceedings of the Conference on Empirical Processing, in Natural Language, 2016, pp. 606–615.

[4] D. Ma, S. Li, X. Zhang, and H. Wang, “Interactive attention networks for aspect-level sentiment classification,” in Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017, pp. 4068–4074.

[5] S. Ahmad and U. Chandra, “Sentiment Index and Sentiment Score of Mobile Phones,” no. October, 2019.

[6] S. O. Reilly, “Decoding Sentimental Tweets,” 2015.

[7] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” J. Comput. Sci., vol. 2, no. 1, pp. 1–8, 2011.

[8] G. A. Buntoro, “Sentiment Analysis of Candidates for Governor of DKI Jakarta 2017 on Twitter (*Analisis Sentimen Calon Gubernur DKI Jakarta 2017 di* Twitter),” Integer J. Maret, vol. 1, no. 1, pp. 32–41, 2017.

[9] N. C. Wirawan and P. P. Adikara, “Sentiment Analysis with Query Expansion on M-Banking Application Review using Fuzzy K-Nearest Neighbor Method (*Analisis Sentimen Dengan Query Expansion Pada Review Aplikasi M- Banking Menggunakan Metode* Fuzzy K-Nearest Neighbor (Fuzzy k-NN)),” *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. 1, pp. 362–368, 2018.

[10] K. Pasupa, T. Seneewong, and N. Ayutthaya, “Thai sentiment analysis with deep learning techniques : A comparative study based on word embedding , POS-tag , and sentic features,” Sustain. Cities Soc., vol. 50, no. May, p. 101615, 2019.

[11] K. D. Adhine Salsabila, A. Ghiffarie, R. P. Baistama, M. I. Variadi, and M. D. Rhajendra, "Sentiment Analysis of The Body Shop Tea Tree Oil Products (*Analisis Sentimen Terhadap Produk* The Body Shop Tea Tree Oil)," J. Teknol. dan Manaj. Inform., vol. 5, no. 2, 2019.

[12] R. Mahendrajaya, G. A. Buntoro, and M. B. Setyawan, "Gopay User Sentiment Analysis Using the Lexicon Based Method and Support Vector Machine (*Analisis Sentimen Pengguna Gopay Menggunakan Metode* Lexicon Based d*an* Support Vector Machine)," Komputek, vol. 3, no. 2, p. 52, 2019.

[13] V. L. Nguyen, D. Kim, V. P. Ho, and Y. Lim, "A New Recognition Method for Visualizing Music Emotion," in International Journal of Electrical and Computer Engineering (IJECE), 2017, vol. 7, no. 3, pp. 1246–1254.

[14] F. Zablith and I. H. Osman, "ReviewModus : Text classification and sentiment prediction of unstructured reviews using a hybrid combination of machine learning and evaluation models," Applied Mathematical Modelling, vol. 71, no. 2019, pp. 569–583, 2020.

[15] I. Imamah, H. Husni, E. M. Rohman, I. O. Suzanti, and F. A. Mufarroha, "Text mining and Support Vector Machine for Sentiment Analysis of tourist Reviews in Bangkalan Regency," J. Phys. Conf. Ser., vol. 1477, no. 2, pp. 0–6, 2020.

[14] O. Somantri and D. Dairoh, "Sentiment Analysis of Tegal City Tourism Destination Assessment Based on Text Mining (*Analisis Sentimen Penilaian Tempat Tujuan Wisata Kota Tegal Berbasis Text Mining*)," J. Edukasi dan Penelit. Inform., vol. 5, no. 2, p. 191, 2019.

[15] H. Wu and X. Gu, "Reducing over-weighting in supervised term weighting for sentiment analysis," COLING 2014 - 25th Int. Conf. Comput. Linguist. Proc. COLING 2014 Tech. Pap., pp. 1322–1330, 2014.

[16] H. Hamdan, P. Bellot, and F. Bechet, "Lsislif: CRF and Logistic Regression for Opinion Target Extraction and Sentiment Polarity Analysis," no. SemEval, pp. 753–758, 2015.