Sentiment Analysis of Public Reaction to COVID19 in Twitter Media using Naïve Bayes Classifier

Nur Iksan
Faculty of Engineering
Universitas Negeri Semarang
Semarang, Indonesia
nur.iksan@mail.unnes.ac.id

Djoko Adi Widodo
Faculty of Engineering
Universitas Negeri Semarang
Semarang, Indonesia
djokoadiwidodo@mail.unnes.ac.id

Budi Sunarko
Faculty of Engineering
Universitas Negeri Semarang
Semarang, Indonesia
budi.sunarko@mail.unnes.ac.id

Erika Devi Udayanti
Faculty of Computer Science
Universitas Dian Nuswantoro
Semarang, Indonesia
erikadevi@dsn.dinus.ac.id

Etika Kartikadharma
Faculty of Computer Science
Universitas Dian Nuswantoro
Semarang, Indonesia
etika.kartikadarma@dsn.dinus.ac.id

***Abstract*— Currently, the world's attention was focused on the disease outbreak, namely the corona virus (COVID19). World Health Organization (WHO) declare that this virus was a global pandemic in all countries. The various impacts that arise due to this virus cover various fields, namely health, social, political, religious, economic to resilience and security. Some of the services currently used were still focused on the health sector, namely in the form of treatment and information services related to the development of the spread of the virus. This research will develop a service that was used to identify social impacts in the community through observing community activities on social media, namely Twitter, in the form of an analysis of the public's reaction to COVID19. Through this Twitter, a data acquisition process will be carried out to obtain data related to COVID19 which will then be carried out a sentiment analysis using the Naïve Bayes method so that the results of the public reaction sentiment will be obtained. The experimental result shows that prediction accuracy was 0,86. Furthermore, the results of the Recall was 0,687, the precision was 0,827 and the F-Score was 0.749**

Keywords—Sentiment Analysis, Twitter, Covid19, Naïve Bayes

I. INTRODUCTION

The Corona virus (COVID-19) has been declared a pandemic outbreak by the World Health Organization (WHO), which spreads in all countries globally. In Indonesia, the spread of this virus began to be known in early 2020 [1]. Several information services related to the spread of this virus have been provided by the government [2] to increase public awareness so as to prevent the spread of this virus. Furthermore, another service in the form of a pandemic disease monitoring system was also provided by Google through the Google Flu Trends (GFT) application. Furthermore, another service in the form of a pandemic disease monitoring system was also provided by Google through the Google Flu Trends (GFT) application. This application uses the BDA approach and can estimate the level of Influenza virus infection based on user search patterns [3]. However, these services [1], [2] require a relatively long data collection time and the GFT application [3] still has a low level of accuracy in estimating [4].

In addition to having an impact on health, other impacts caused by the spread of this virus include social, political, economic, resilience and security which cannot be identified from previously provided services [1], [2]. This research will identify the social impacts on society as a result of this virus. This impact can be known through direct observation or through social media on community activities. Through social media, community activities can be analysed so that the social impacts caused by COVID19 can be identified. This social media will become a social sensor to obtain data from social media such as Twitter, Facebook, Instagram and others. Data from social media contains information related to activities carried out by the community and information when encountering certain events.

This research will develop a service that was used to identify social impacts in the community through observing community activities on social media, namely Twitter, in the form of an analysis of the public's reaction to COVID19. Through this Twitter, a data acquisition process will be carried out to obtain data related to COVID19 which will then be carried out a sentiment analysis using the Naïve Bayes method so that the results of the public reaction sentiment will be obtained.

In this research, the Sentiment Analysis method will be developed using Naïve Bayes as a Learning model to identify social impacts in the community. This paper consists of five chapters, including chapter one discusses the background, chapter two discusses related research, chapter three discusses Sentiment Analysis method methods for public reaction identification, chapter four discusses the analysis of experimental results, chapter five contains conclusions

II. RELATED WORKS

Social media platforms can be used to develop pandemic disease monitoring systems and have the potential to increase the lag time available to pre-existing services [1], [2], [3]. The data available on social media was very large and varied and was generally in the form of expressions of users' feelings and health conditions. This social media platform was also referred to as social censorship because it gets / obtains data related to social activities from the community as users. One of the most popular social media platforms was Twitter. Several studies have developed the Twitter platform as a social sensor in a health monitoring system [5], [6] [7] [8]. In addition, it was also used in disaster monitoring systems, marketing, politics and so on.

Several studies have been carried out in developing sentiment analysis methods related to the management of pandemic diseases. Researcher [9] conducted an analysis of Twitter content during the H1N1 outbreak and measured sentiment in a qualitative categorical way. Researcher [10] developed an Epidemic Sentiment Monitoring System (ESMOS) application to analyse disease sentiment and measure the Measure of Concern (MOC) using the number of negative tweets per day expressed by users. Researchers [11]

978-1-6654-4181-0/21/$31.00 ©2021 IEEE
July 14-16th 2021

[12] used a Degree of Concern (DOC) which was calculated and visualized with a personal Tweet. Researcher [13] used multi modal data obtained from social media to be further analysed and produced recommendations related to personal health. Researcher [6] used the lexicon method with tweet content in the form of tweet geolocation, time, and user ID, Uniform Resource Locators (URLs) and emojis.

This research uses social media Twitter to obtain data related to COVID19 in Indonesia for further analysis. The monitoring system was more directed at identifying people's reactions to the COVID19 pandemic disease issue. The data source was obtained from Twitter via the Twitter API.

III. Sentiment Analysis Method

Sentiment analysis aims to analyze social media data into information contained in the text. Sentiment analysis was also called Opinion Mining which analyzes a person's opinion or opinion, emotions, behavior on a topic or issue. Sentiment analysis from social media, especially Twitter, was the subject of research that was mostly carried out by researchers to identify opinions that develop in society. These opinions can be in the form of public reactions to a service, product, health, marketing, and others. There were two main methods used in sentiment analysis, namely the Lexicon method and the Machine Learning method.

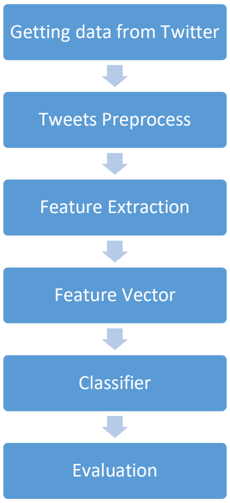


Fig. 1. Stages of Twitter Data Sentiment analysis

In the Lexicon method, each word collection of words has a score indicating positive, negative, neutral and objective characteristics of the text. In certain parts of the text, the score and the highest score were combined to give the overall polarity of a text. The Lexicon method uses a dictionary or dictionary to score a word, then several words will be paired based on their polarity values. The dictionary approach uses word relations such as synonyms, antonyms, hypernym and hyponyms in a collection of words to obtain opinion words. Besides the dictionary, another approach used in the Lexicon method was the Corpus. The Corpus approach uses a collection of opinion words as seeds and syntactic patterns. Sentiment analysis on Corpus can be done statistically and semantically. A statistical way by finding words that appear together in the Corpus. If most of the words that appear were positive text then the polarity was positive and vice versa. The semantic way was done by calculating the sentiment values using the principle of the similarity between words

Overall, the stages in sentiment analysis include 5 main stages as shown in Figure 1, namely Data Collection, Pre-processing, Feature Extraction, Classification and validation. In the Feature Extraction stage, sentiment analysis that uses a statistical approach utilizes statistical calculations such as the frequency of word occurrences which was usually called the Term Frequency (TF) of the entire document or Inverse Document Frequency (TF-IDF). In addition, the Feature Extraction stage uses semantic approaches to lexicons. The pre-processing stage aims to prepare documents so that they can be processed at a later stage. At this stage, the process that can be carried out was tokenizing (cutting words), formalization (adjusting to the KBBI standard), translate, POS tagging, filtering (making stop words) and stemming (changing to basic word forms). Furthermore, the feature extraction stage was used to retrieve features that will be used in the classification stage

The preprocessing stage aims to prepare documents so that they can be processed at a later stage. Pre-processing stage [9] were divided two part functionality, including common functionality for tasks related to the existing text in the tweet and specific functionality for tasks related to the characteristics of the tweet, as shown in Figure 2.

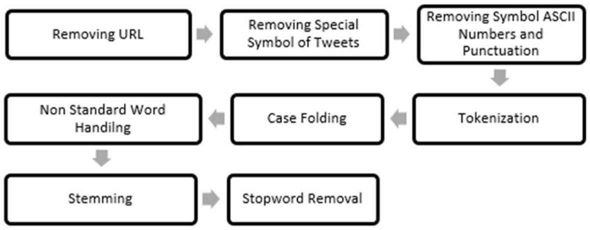

Removing URL
Removing Special Symbol of Tweets
Removing Symbol ASCII Numbers and Punctuation
Tokenization
Case Folding
Non Standard Word Handilng
Stemming
Stopword Removal

Fig. 2. Pre-processing stages

Common functionality consists of several tasks, ie:

1) Removing URL: The tweet contains a link to various sources. we can remove the URL because it was not used in this process. We can convert these emoticons in the form of word.

2) Removing special symbol of tweets: Some special symbols on twitter such as hashtag (#), retweet (RT) and username (mention) username can be removed at this task.

3) Tokenization: This task will divide the sentence into several parts or tokens which can be in the form of words and phrases.

4) Case Folding: changes the uppercase of a word to lowercase.

5) Non Standard Word Handling: this task will handle non-standard word usage such as abbreviations, miss spelling, slang words, lengthening words. As an example of a non-standard word in Indonesian, namely 'horeeee' from the standard word 'hore' which means 'hurray' in English.

6) Stemming: this task will transform the word into its basic word form by removing all affixes which consist of prefixes, infixes, suffixes and confixes. Sastrawi library [10] was used in this task.

7) Stopword Removal : this taks will remove common words does not have a significant effect in the sentence. A list of stopword words was also found in the Sastrawi library for text processing in Indonesian. Some of the words that were included in the stopword list include 'and', 'then', 'with' and so on.

Furthermore, the feature extraction stage using TF-IDF method and produce term-document matrix i.e. TF and IDF matrix that will be used in the classification stage. The weighting of each TF-IDF matrix uses the following formula:

$$tf.idf_{t.d} = tf_{td} \times idf_t \quad (1)$$

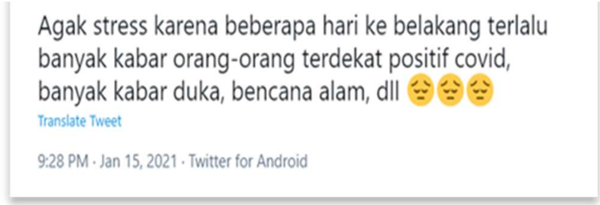
$$idf_t = log \frac{N}{df_t} \quad (2)$$

Where $tf.idf_{t.d}$ denotes the weight of each word from term *t* contained in the document *d*. tf_{td} denotes the weight term *t* contained in the document *d*. idf_t denotes inverse the document frequency contained in term *t*. *N* denotes the number of all documents. df_t denotes the number of documents containing term *t*. *t* denotes term or word.

The accuracy value was obtained by comparing the amount of data from the classification results with the total number of data. The higher the accuracy value obtained, the better the classification results in the method used. To see whether or not there was data deviation, the Recall, Precision and F-Score values were also calculated. Recall was obtained by comparing the number of classified data relevant to the total data that was considered relevant. While precision was obtained by comparing the number of relevant classification results data with the total amount of data in a particular class. The F-score was the result of the average value obtained from precision and recall. After measuring the performance of the methods used, the next experiment was to classify the public reactions, namely Angry, Sad, Afraid, Love, Happy.

IV. Result and Discussions

The data collection process was carried out via twitter with the keyword “covid19 OR corona” in which are captured in realtime and can be arranged based on a specific date and using Indonesian. The input data were taken from Twitter using the Search API. The Twitter API allows users to get tweets with language filters. The example of Twitter tweet data can be seen in Figure 3.



Agak stress karena beberapa hari ke belakang terlalu banyak kabar orang-orang terdekat positif covid, banyak kabar duka, bencana alam, dll 😔😔😔

9:28 PM · Jan 15, 2021 · Twitter for Android

Fig. 3. The Crawling Example of Twitter Tweet Data

Figure 3 was the result of the capture of the Tweeter user's tweet text posted during the COVID19 pandemic. The text of the tweet was written by a user from Indonesia in Indonesian. If translated into English the text of the tweet becomes as follows:

> *“It's a bit stressful because the past few days there has been too much news from those closest to being positive for covid, a lot of sad news, natural disasters, etc.”*

The tweet text will then be processed in the sentiment analysis stage as shown in Figure 1.

In the pre-processing process, several tasks include removing URL, converting emoticons, removing special symbol of tweets, removing symbol ASCII numbers and punctuation, tokenization, case folding, non-standard word handling, stemming, stop word removal. Figure 3 shows the results of the pre-processing process. Based on this figure, the preprocessing process is carried out on the Twitter text with several preprocessing stages, as mentioned in section 3. For example, the result of crawling and preprocessing tweet text is shown in Table I. In this table, some words from the original tweet have been removed through the function of stop word, punctuation and stemming. Then cutting words of tweet preprocessed result by the function of tokenization.

TABLE I. Tweet Dataset Preprocessing

Preprocessing Stages	**Original Tweet**	**Tweet in English**
Tweet	*“Agak strees karena beberapa hari ke belakang terlalu banyak kabar orang-orang terdekat positif covid, banyak kabar duka, bencana alam, dll* 😔 😔 😔 *"*	*“It's a bit stressful because the past few days there has been too much news from those closest to being positive for covid, a lot of sad news, natural disasters, etc.* 😔 😔 😔 *”*
Preprocesing Result	*“stress kabar orang orang dekat positif covid kabar duka bencana alam“*	*"stress news of close people positive for covid, sad news of natural disasters"*
Tokenization	“[stress] [kabar] [orang] [orang] [dekat] [positif] [covid] [kabar] [duka] [bencana] [alam]”	“[stress] [news] [people] [people] [nearby] [positive] [covid] [news] [grief] [disaster] [natural]”

In this research, there were five classes, namely Angry, Sad, Afraid, Love, Happy. The Naive Bayes classification was done to get a model that can predict the class of new tweets. To determine the training model, 3000 datasets were used, each of which was equipped with a label or class. Furthermore, the data was used in the classification process on new tweets with a total of 500 tweets. In this classification process, the training model that has been prepared was used to predict new tweets to determine each class. The training process on the Naive Bayes classification begins by calculating the probability of a tweet appearing based on its class. The results of this classification were then compared with the results of manual classification so that the performance test values were obtained as shown in Figure 4. The experimental scenario that has been done was calculating the results of accuracy, recall, precision, and f-score on the data with the distribution of each class. The experimental result shows that prediction accuracy was 0,86. Furthermore, the results of the Recall was 0,687, the precision was 0,827 and the F-Score was 0.749.

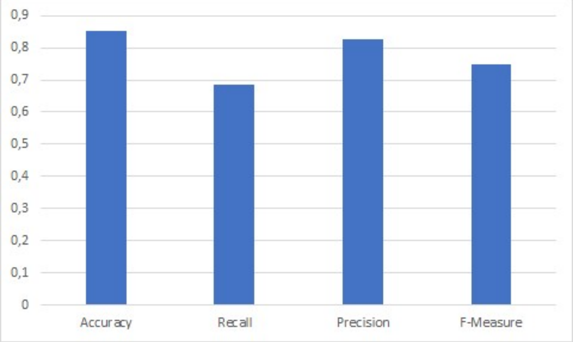


Fig. 4. The result of the Accuracy, Recall, Precision and F-Score measurements

Furthermore, the results of the training model were used to perform real-time classification on the covid 19 monitoring system with the same keywords, namely "covid19 OR corona" using Indonesian. The visual result was shown in the form of a word cloud from the words that appear in each class in realtime. Figure 5 below shows dashboard of covid19 monitoring system namely a word cloud of popular topics discussed and classified in each class.

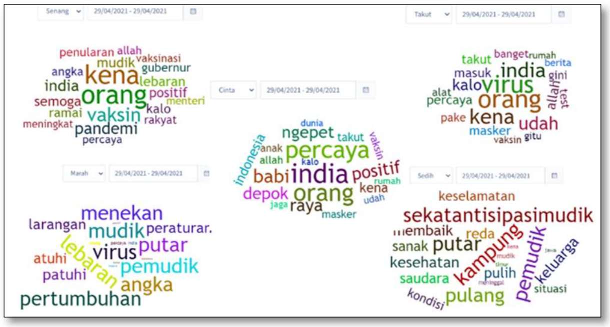

Senang
29/04/2021 - 29/04/2021
penularan allah mudik vaksinasi gubernur angka kena lebaran india orang positif menteri semoga ramai vaksin kalo rakyat meningkat pandemi percaya
Cinta
29/04/2021 - 29/04/2021
dunia ngepet takut vaksin indonesia anak allah percaya kalo babi india positif rumah kena depok orang udah jaga raya masker
Takut
29/04/2021 - 29/04/2021
takut banget rumah berita masuk india gini alat kalo virus allah test percaya orang pake kena udah masker vaksin gitu
Marah
29/04/2021 - 29/04/2021
menekan larangan mudik peraturan lebaran virus putar atuhi pemudik patuhi angka pertumbuhan
Sedih
29/04/2021 - 29/04/2021
keselamatan sekatantisipasimudik membaik reda sanak putar kampung pemudik keluarga kesehatan pulih saudara situasi kondisi pulang

Fig. 5. Dashboard of Covid19 Monitoring System: Word Cloud of popular topics discussed

This Word cloud was taken on April 29, 2021. Many things have become the topic of discussion from Indonesian user at this time that discuss issues inside and outside the country. Through this Word Cloud, popular topics that are discussed from each community reaction will be known, which are shown in Table II.

TABLE II. WORD CLOUD OF POPULAR TOPICS DISCUSSED FROM INDONESIAN USERS

Label Class	**Original Word CLoud**	**Word Cloud in English**
Happy : "*Senang*"	*"vaksin, mudik, lebaran, allah, gubernur,…"*	vaccines, going home, Eid, Allah, governor,...
Angry: "*Marah*"	*"larangan, peraturan, putar, virus, mudik,…"*	Prohibitions, regulations, play, viruses, homecoming,...
Sad: "*Sedih*"	*"Sekatan, putar, kampung, pemudik, situasi, pulang,…"*	Block, turn, village, homecoming, situation, go home,...
Afraid : "*Takut*"	"*india, virus, berita, masker, kena, test,…"*	India, virus, news, masks, got hit, test,...
Love : "*Cinta*"	"*vaksin, india, raya, masker, jaga, percaya,…"*	Vaccines, India, Raya, masks, guard, believe,...

V. CONCLUSION

This paper presents a process for identifying public reactions through sentiment analysis on Twitter messages in Indonesian. The pre-processing stage starts from the process of removing URLs, removing special symbols, removing ASCII symbols, punctuation, tokenization, case folding, stopword and stemming. The extraction process uses TF-IDF and generates a feature vector which was then used for the identification or sentiment analysis of public reaction. The experimental result shows that prediction accuracy was 0,86. Furthermore, the results of the Recall was 0,687, the precision was 0,827 and the F-Score was 0.749.

ACKNOWLEDGMENT

We would like to thank to The Directorate of Research and Community Service – Ministry of Research, Technology and Higher Education (DRPM – Kemenristekdikti); and Semarang State University (UNNES) for their assistance in facilitating research.

REFERENCES

[1] GTPPCovid19, "*Map of the Distribution of the COVID-19 Virus in Indonesia*," July, 2020.

[2] PemprovJateng, "Distribution of COVID-19 Cases in Central Java," July 2020.

[3] A. F. Dugas, M. Jalalpour, Y. Gel, S. Levin, F. Torcaso, T. Igusa and R. E. Rothman, "Influenza Forecasting with Google Flu Trends," PLOS ONE, vol. 8, no. 2, 2013.

[4] K. Byrd, A. Mansurov and O. Baysal, "Mining Twitter Data for Influenza Detection and Surveillance," in IEEE/ACM International Workshop on Software Engineering in Healthcare Systems (SEHS), USA, 2016.

[5] S. E. Jordan, S. E. Hovet, I. C.-H. Fung, H. Liang, K.-W. Fu and Z. T. H. Tse, "Using Twitter for Public Health Surveillance from Monitoring and Prediction to Public Response," MDPI, vol. 4, no. 1, 2018.

[6] B. Alkouz, Z. A. Aghbari and J. H. Abawajy, "Tweetluenza: Predicting flu trends from twitter data," Big Data Mining and Analytics, vol. 2, no. 4, 2019.

[7] S. Sidana, S. Amer-Yahia, M. Clausel, M. Rebai, S. T. Mai and M.-R. Amini, "Health Monitoring on Social Media over Time," IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 8, 2018.

[8] K. Lee, A. Agrawal and A. Choudhary, "Forecasting Influenza Levels Using Real-Time Social Media Streams," in IEEE International Conference on Healthcare Informatics (ICHI), USA, 2017.

[9] G. E. Cynthia Chew, "Pandemics in the age of twitter: content analysis of Tweets during the 2009 H1N1 outbreak," PLoS One, 2010.

[10] X. Ji, S. A. Chun and J. Geller, "Knowledge-Based Tweet Classification for Disease Sentiment Monitoring," in Sentiment Analysis and Ontology Engineering, USA, Springer, 2016, pp. 425-454.

[11] S. M. Mohamma, S. Kiritchenko and X. Zhu, "Detecting public sentiment over PM2.5 pollution hazards through analysis of Chinese microblog," in international workshop on Semantic Evaluation Exercises (SemEval-2013), USA, 2013.

[12] X. ji, S. A. Chun and J. Geller, "Monitoring public health concerns using twitter sentiment classifications," in IEEE International Conference on Healthcare Informatics, 2013.

[13] X. Zhou, W. Liang, K. I.-K. Wang and S. Shimizu, "Multi-Modality Behavioral Influence Analysis for Personalized Recommendations in Health Social Media Environment," IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, vol. 6, no. 5, 2019.

[14] A. F. Hidayatullah and M. R. Maarif, "Pre-processing Tasks in Indonesian Twitter Messages," in International Conference on Computing and Applied Informatics, Medan, Indonesia, 2016.