# BUKTI KORESPONDENSI ARTIKEL PADA JURNAL INTERNASIONAL BEREPUTASI

**PENGUSUL**

Dr. Djoko Adi Widodo, MT / NIDN 0027095912

**UNIVERSITAS NEGERI SEMARANG**

Yth. Penilai
Pada Usulan PAK

Bersama dengan surat ini, saya bermaksud menyertakan bukti bukti korespondensi proses artikel pada Jurnal Internasional dengan judul "Sentiment Analysis of Twitter Media for Public Reaction Identification on COVID-19 Monitoring System using Hybrid Feature Extraction Method" yang dimuat pada *International Journal of Intelligent Systems and Applications in Engineering-IJISAE* Volume 11, Issue 1, Published Online 16 Januari 2023.

Adapun susunan kronologi bukti korespondensi terdiri dari beberapa poin, pada tabel di bawah ini:

| No | Tanggal | Aktivitas |
|---|---|---|
| 1. | 2 Agustus 2022 | Mempersiapkan final draft jurnal. |
| 2. | 4 Agustus 2022 | Menyesuaikan draft jurnal dengan template pada *International Journal of Intelligent Systems and Applications in Engineering-IJISAE*<br>(link template jurnal: https://ijisae.org/index.php/IJISAE/about) |
| 3. | 14 Agustus 2022 | Pembuatan akun submission pada sistem IJISAE https://www.manuscriptsubmission.net/ ijisae/index.php/submission/about/submissions |
| 4. | 14 Agustus 2022 | Submit manuskrip pertama (ID Paper: 158) |
| 5. | 19 Agustus 2022 | Pemberitahuan via email bahwa naskah ditolak dengan alasan persentase plagiarism Turnitin melebihi standart IJISAE. |
| 6. | 20 Agustus 2022 | *Proof-reading* naskah dan melakukan pengecekan Turnitin, revisi untuk penurunan persentase plagiarism. |
| 7. | 5 Desember 2022 | Submit ulang manuskrip dan *cover letter* sebagai berkas pendukung (ID Paper: 671). |
| 8. | 5 Desember 2022 | Pemberitahuan via email bahwa manuskrip telah berhasil di submit dan telah diterima untuk direview. |
| 9. | 11 Desember 2022 | Pemberitahuan via email bahwa manuskrip telah diterima untuk publikasi, permintaan pembayaran *article publication* dan perbaikan manuskrip sesuai template. |
| 10. | 15 Desember 2022 | Mengirim manuskrip sesuai template, bukti pembayaran dan *copyright declaration* |
| 11. | 21 Desember 2022 | Pemberitahuan registrasi untuk publikasi diterima |

| No | Tanggal | Aktivitas |
|---|---|---|
| 12. | 16 Januari 2023 | Pemberitahuan bahwa manuskrip telah dipublikasi online.https://ijisae.org/index.php /IJISAE/article/view/2447 |

Demikian, agar dapat menjadi periksa.
Terima kasih.

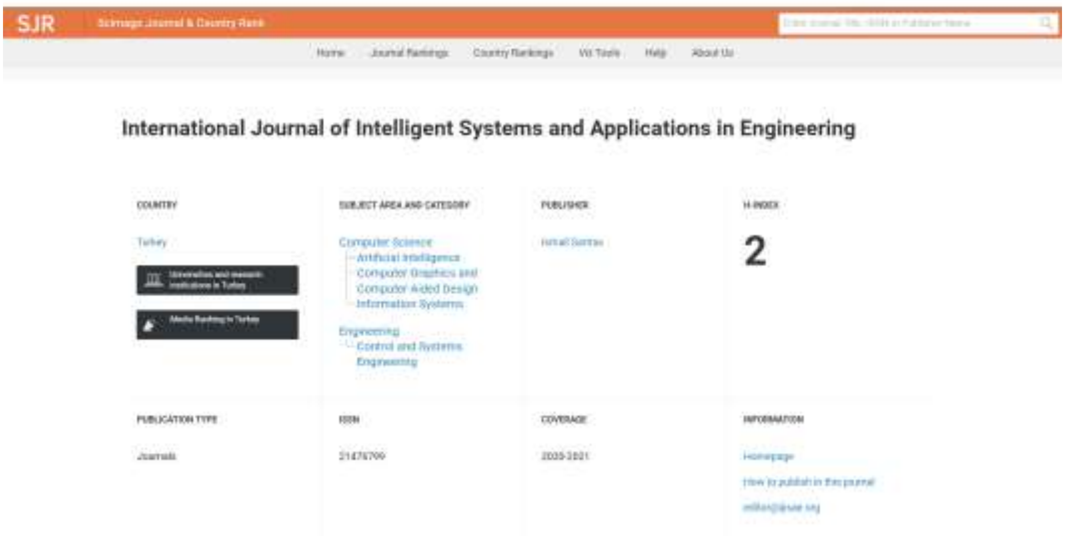
Semarang, 7 Agustus 2023
Hormat saya,

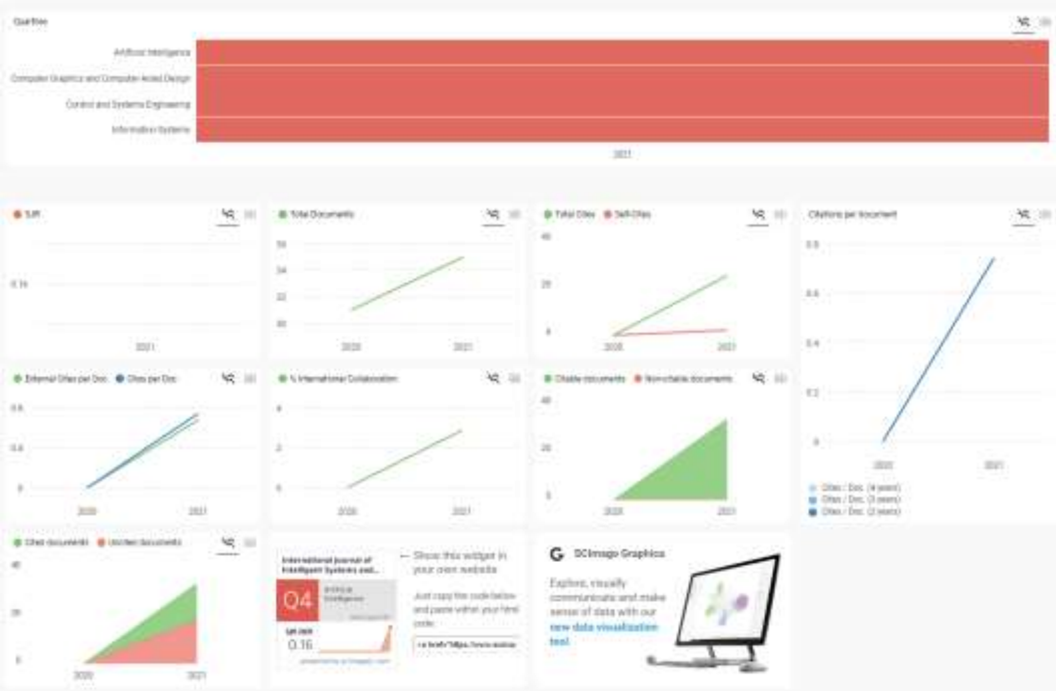
Djoko Adi Widodo

## KRONOLOGI KORESPONDENSI PUBLIKASI ARTIKEL PADA JURNAL INTERNASIONAL BEREPUTASI DAN BERFAKTOR DAMPAK

| | |
|---|---|
| Judul | : Sentiment Analysis of Twitter Media for Public Reaction Identification on COVID-19 Monitoring System using Hybrid Feature Extraction Method |
| Jurnal | : *International Journal of Intelligent Systems and Applications in Engineering-IJISAE* |
| Volume | : 11 |
| Nomor | : 2 |
| Tanggal Publikasi | : 16 Januari 2023 |
| ISSN | : 2147-6799 |
| Halaman | : 92-99 |
| Penerbit | : Ismail Saritas |
| SJR Journal | : 0.157 |
| Quartile | : Q4 |
| Cite Score | : 4.2 |
| SNIP | : 0.122 |
| Penulis | : Djoko Adi Widodo, Nur Iksan, Budi Sunarko |

Bukti indexing jurnal:




Artificial Intelligence
Computer Graphics and Computer-Aided Design
Control and Systems Engineering
Information Systems
SJR
Total Documents
Total Cites
Self-Cites
Citations per document
% International Collaboration
Q4
SJR 2021
0.16
SCImago Graphica


International Journal of Intelligent Systems and Applications in Engineering
IJISAE
Register
Login
ISSN : 2147-6799
HOME
CURRENT
ARCHIVES
ANNOUNCEMENTS
ABOUT
Search
About the Journal
Information for Authors:
We are pleased to inform that we are now collaborating with **Elsevier Digital Commons** for much better visibility of journal. Further authors will be able to observe their citations, metric like PlumX from journal website itself. **IJISAE** will be in transition from **OJS** to **Digital Commons framework** in next few months so if their is any queries or delays contact directly on ***editor@ijisae.org***
International Journal of Intelligent Systems and Applications in Engineering (IJISAE) is an international and interdisciplinary journal for both invited and contributed peer reviewed articles that intelligent systems and applications in engineering at all levels. The journal publishes a broad range of papers covering theory and practice in order to facilitate future efforts of individuals and groups involved in the field. **IJISAE**, a peer-reviewed double-blind refereed journal, publishes original papers featuring innovative and practical technologies related to the design and development of intelligent systems in engineering. Its coverage also includes papers on intelligent systems applications in areas such as nanotechnology, renewable energy, medicine engineering, Aeronautics and Astronautics, mechatronics, industrial manufacturing, bioengineering, agriculture, services, intelligence based automation and appliances, medical robots and robotic rehabilitations, space exploration and etc.
As an Open Access Journal, IJISAE devotes itself to promoting scholarship in intelligent systems and applications in all fields of engineering and to speeding up the publication cycle thereof. Researchers worldwide will have full access to all the articles published online and be able to download them with zero subscription fees. Moreover, the influence of your research will rapidly expand once you become an Open Access (OA) author, because an OA article has more chances to be used and cited than does one that plods through the subscription barriers of traditional publishing model.
**IJISAE (ISSN: 2147-6799)** indexed by SCOPUS, TR Index, IndexCopernicus, Global Impact Factor, Cosmos, Google Scholar, JournalTocs, IdeaKOnline, OAJI, ResearchGate, ESJI, Crossref and ROAD.
ANNOUNCEMENTS
Information for Authors
March 6, 2023
Information for Authors:
We are pleased to inform that we are now collaborating with **Elsevier Digital Commons** for much better visibility of journal. Further authors will be able to observe their citations, metric like PlumX from journal website itself. **IJISAE** will be in transition from **OJS** to **Digital Commons framework** in next few months so if their is any queries or delays contact directly on ***editor@ijisae.org***
NON SOLUS
ELSEVIER


Scopus
Author Search
Sources
Source details
International Journal of Intelligent Systems and Applications in Engineering
Scopus coverage years: from 2020 to 2022
Publisher: Ismail Saritas
E-ISSN: 2147-6799
Subject area:
Source type: Journal
CiteScore 2021
0.4
0.157
SNIP 2021
0.122
CiteScore
CiteScore rank & trend
Scopus content coverage
Improved CiteScore methodology
CiteScore 2021
0.4 = 34 Citations 2018 - 2021 / 66 Documents 2018 - 2021
CiteScoreTracker 2022
1.2 = 305 Citations to date / 242 Documents to date

# RINCIAN KRONOLOGI KORESPONDENSI PUBLIKASI ARTIKEL PADA JURNAL INTERNASIONAL BEREPUTASI

1. **Mempersiapkan final draft jurnal.**

# Tweet Sentiment Classifier on COVID19 Monitoring System for Public Reaction Identification Toward COVID-19

**Djoko Adi Widodo[1], Nur Iksan*[2], Budi Sunarko[3], Erika Devi Udayanti[4]**

Submitted: xx/xx/202x Accepted : xx/xx/201x DOI: 10.1039/b000000x

***Abstract:*** Currently, the Corona virus (COVID19) becomes the disease outbreak causing serious concern. The virus has been declared by the WHO as a pandemic spreading all over countries globally, including Indonesia. The spread of this virus has been prevented by implementing social distancing, quarantine and work from home, mandatory use of mask, hand washing and others. However, this method is not yet effective because not all people are aware and disciplined to apply it in their daily lives, so the spread of this virus still has potential as a threat. This virus does not only affect health, but also affects social, political, religious, economic and resilience aspects found through direct observation to the public or through social media. Community activities can be analyzed though social media so that the impact caused by COVID19 can be known, especially the socio-economic impact on society. This paper uses Twitter social media to obtain data related to COVID19 in Indonesia for further analysis. The analysis of Twitter data was done to identify trends related to the spread of Corona virus and social impact in society through public reactions to the COVID19 issue using the data mining. This research developed a sentiment analysis model using TF-IDF and lexical methods to identify public reactions due to COVID19. Data sources were obtained from Twitter accessed via the Twitter API. Features used matrix score from TF-IDF and lexicon-based method. Prediction accuracy measurement result shows our proposed method is more accurate than the other method which is 0,854. Furthermore, the results of the Recall, Precission and F-Score measurements shows the ME method has poor performance while AFBS and FBS have better performance with superior AFBS.

***Keywords:*** *Tweeter Analysis, Sentiment Classifier, TF-IDF, Lexicon Based, Covid-19*

## 1. Introduction

Outbreaks of disease that occur in the world are serious concern to both the wider community and health agencies such as WHO (World Health Organization). Currently, the outbreak of disease causing a serious concern is the Corona virus (COVID19). This virus has been declared by WHO as a pandemic spreading all over countries globally. In Indonesia, the spread of this virus started to appear in early January 2020. Based on data from the Covid-19 handling team in Indonesia, there were almost 600,000 people were infected by Covid-19 and around 18,000 people died from Covid-19 [1].

Several methods have been done to prevent the spread of this

pandemic that continues to spread and threat all countries. One of the main goals of giving vaccines is to provide immunity against COVID19 infections and then can further encourage the formation of herd immunity. To increase public awareness in efforts to prevent the spread of this virus, the government and several institutions have provided information services for the spread of this Corona virus. This information is very important for the community in order to increase their vigilance in carrying out their activities.

Several other impacts made due to this virus are social, political, religious, economic to resilience and security impacts. These impacts cannot be identified from previous information services [3] However, direct observation or social media on community activities can be taken to identify these impacts. For example

2. **Menyesuaikan draft jurnal dengan template pada *International Journal of Intelligent Systems and Applications in Engineering-IJISAE***

IJISAE

International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

# Sentiment Analysis of Twitter Media for Public Reaction Identification on COVID-19 Monitoring System using Hybrid Feature Extraction Method

**Djoko Adi Widodo*[1], Nur Iksan[1], Budi Sunarko[1]**

Submitted: xx/xx/202x Accepted : xx/xx/201x DOI: 10.1039/b000000x

***Abstract:*** Several strategies were implemented to prevent COVID-19 spread. However, these steps were not effectively implemented in the community due to low public awareness and lack of discipline in daily life and this indicated a potential threat of continuous exposure to the virus. It was also observed that the pandemic greatly affected other areas besides the health sector ranging from the social, political, religious, and economic aspects to the resilience of the people. These can be observed through direct observation of the community or activities of the people on social media, especially in relation to the socio-economic aspect. Therefore, this research was conducted using social media, specifically Twitter, via the Twitter API to obtain data related to COVID-19 pandemic in Indonesia. In this research, a sentiment analysis method was developed in this research to identify public opinion related to the spread of the COVID-19 virus and its social impact on society. This was achieved using the TF-IDF and Lexical methods for feature extraction and Naive Bayes for classification. This research used a dataset obtained through Twitter using the keyword "COVID-19" in Indonesian and manually labelled using 5 categories of reactions i.e., fear, angry, love, sad, and happy. The prediction accuracy values showed that the proposed TFBS method had a higher accuracy value of 0.85 compared to other methods. The performance was also evaluated by calculating the precision, recall, and F-score values for each extraction method used, and the proposed TFBS method was observed to have the highest values while ME had the lowest.

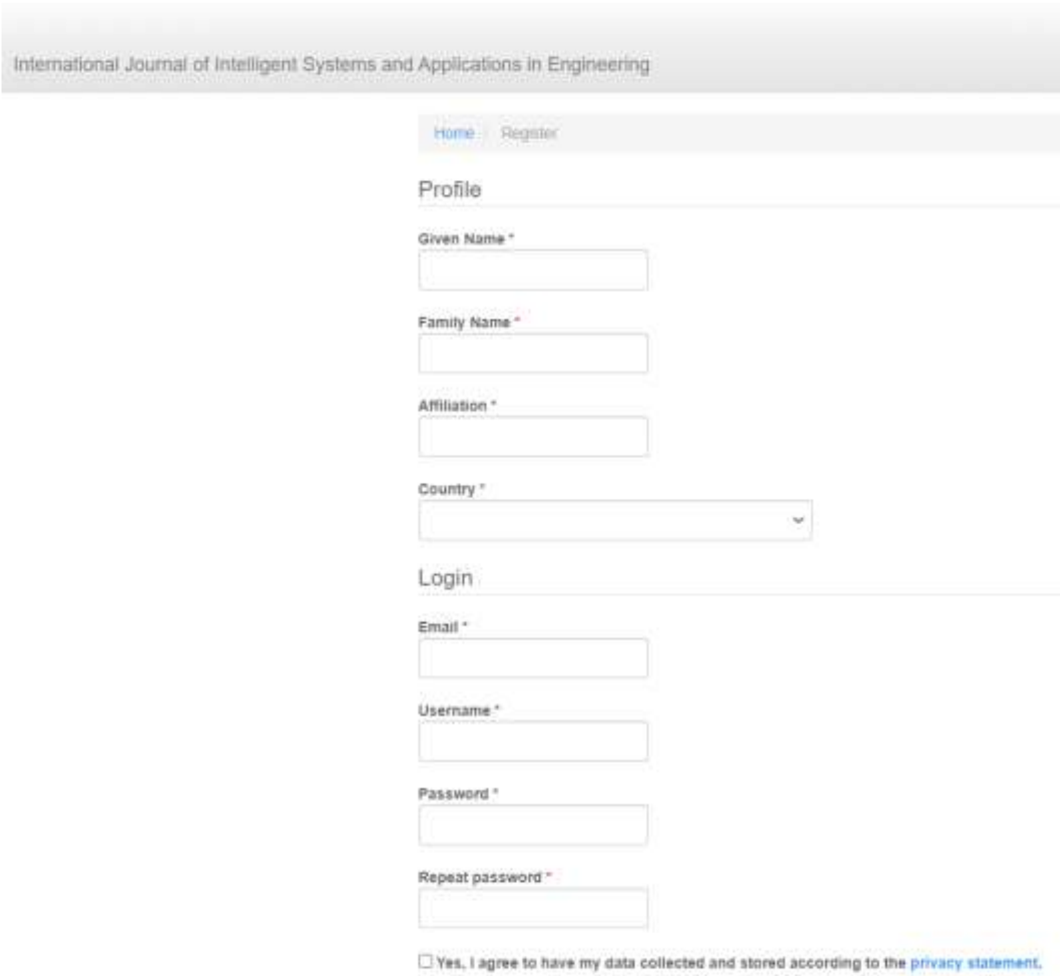
***Keywords:*** *Sentiment Analysis, COVID-19, TF-IDF, Lexicon Based*

## 1. Introduction

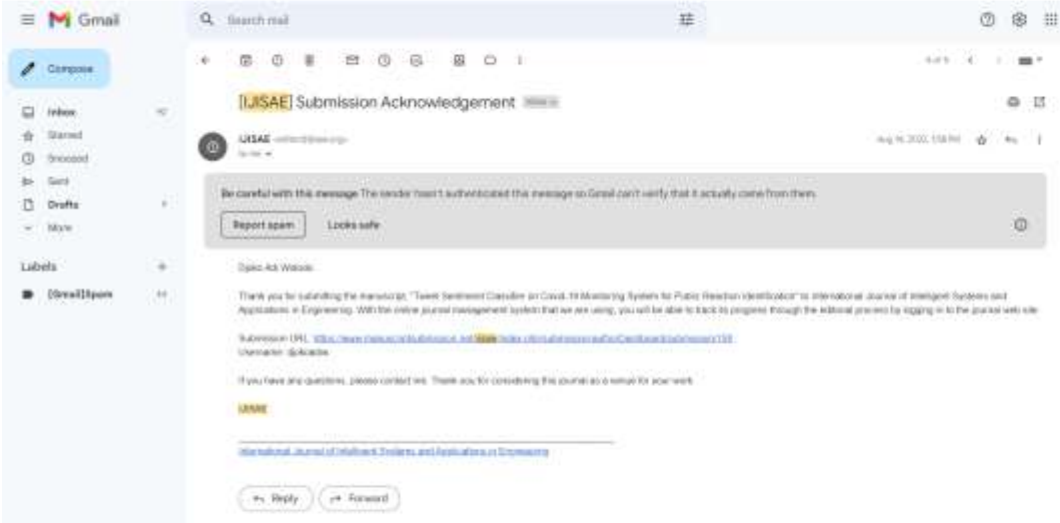
World Health Organization (WHO) declared the COVID-19 virus a pandemic due to its spread to all countries in the world, thereby, becoming a serious concern for the public, government, and world health institutions [1]. The virus started emerging in Indonesia in early January 2020 and the data from the handling team showed an increasing trend with nearly 600,000 people reported to be exposed and approximately 18,000 died. Several strategies were implemented to prevent further spread such as washing hands, wearing masks, quarantine, social distancing, working from home, and others [2]. However, these steps were not effectively implemented in the community due to low public awareness and lack of discipline in daily life and this indicated a potential threat of continuous exposure to the virus. This led the government to introduce another strategy which involved launching a vaccination program for all elements of society to prevent infection.

In addition to the health sector, the pandemic also significantly affected the economic, religious, security, resilience, political, and social sectors [3]. These were not identified directly as the case with the health sector but through the observation of community activities using different media such as Instagram, Facebook, Twitter, and others. These media serve as social sensors for communities through the uploading of ideas, thoughts, and opinions in the forms of text, video, photos, and audio. The contents are focused on activities and conditions of the things the people feel and see in their surrounding environment which are subsequently sha red on social media platforms. Meanwhile, sentiment analysis can be used to identify and predict future conditions using the information [4–6]. A sentiment analysis method was developed in this research to identify public opinion related to the spread of the COVID-19 virus and its social impact on society [7–9]. This was achieved using the TF-IDF and Lexical methods for feature extraction and Naive Bayes for classification.

## 3. Pembuatan akun submission



## 4. Submit manuskrip pertama (ID Paper: 158)



International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

# Tweet Sentiment Classifier on COVID19 Monitoring System for Public Reaction Identification

**Djoko Adi Widodo[1], Nur Iksan*[2], Budi Sunarko[3], Erika Devi Udayanti[4]**

***Abstract:*** Currently, the Corona virus (COVID19) becomes the disease outbreak causing serious concern. The virus has been declared by the WHO as a pandemic spreading all over countries globally, including Indonesia. The spread of this virus has been prevented by implementing social distancing, quarantine and work from home, mandatory use of mask, hand washing and others. However, this method is not yet effective because not all people are aware and disciplined to apply it in their daily lives, so the spread of this virus still has potential as a threat. This virus does not only affect health, but also affects social, political, religious, economic and resilience aspects found through direct observation to the public or through social media. Community activities can be analyzed though social media so that the impact caused by COVID19 can be known, especially the socio-economic impact on society. This paper uses Twitter social media to obtain data related to COVID19 in Indonesia for further analysis. The analysis of Twitter data was done to identify trends related to the spread of Corona virus and social impact in society through public reactions to the COVID19 issue using the data mining. This research developed a sentiment analysis model using TF-IDF and lexical methods to identify public reactions due to COVID19. Data sources were obtained from Twitter accessed via the Twitter API. Features used matrix score from TF-IDF and lexicon-based method. Prediction accuracy measurement result shows our proposed method is more accurate than the other method which is 0,854. Furthermore, the results of the Recall, Precision and F-Score measurements shows the ME method has poor performance while AFBS and FBS have better performance with superior AFBS.

***Keywords:*** *Tweeter Analysis, Sentiment Classifier, TF-IDF, Lexicon Based, Covid-19*

## 1. Introduction

Outbreaks of disease that occur in the world are serious concern to both the wider community and health agencies such as WHO (World Health Organization). Currently, the outbreak of disease causing a serious concern is the Corona virus (COVID19). This virus has been declared by WHO as a pandemic spreading all over countries globally. In Indonesia, the spread of this virus started to appear in early January 2020. Based on data from the Covid-19 handling team in Indonesia, there were almost 600,000 people were infected by Covid-19 and around 18,000 people died from Covid-19 [1].

Several methods have been done to prevent the spread of this virus, including the implementation of social distancing, quarantine work from home, obligatory use of masks, washing hands and others [2]. However, these methods were not effective because not all people are aware and disciplined to adopt those methods in their daily lives so that this virus still has potential as a threat. Furthermore, the government also launches vaccination program for all elements of Indonesia people as an effort to prevent the spread of COVID19. This vaccination program becomes one of effective efforts to overcome the COVID19 pandemic that continues to spread and threat all countries. One of the main goals of giving vaccines is to provide immunity against COVID19 infections and then can further encourage the formation of herd immunity. To increase public awareness in efforts to prevent the spread of this virus, the government and several institutions have provided information services for the spread of this Corona virus. This information is very important for the community in order to increase their vigilance in carrying out their activities.

Several other impacts made due to this virus are social, political, religious, economic to resilience and security impacts. These impacts cannot be identified from previous information services [3]. However, direct observation or social media on community activities can be taken to identify these impacts. For example, community activities can be analyzed through social media to identify social impacts caused by COVID19. This social media becomes a social sensor to obtain data, such as from Twitter, Facebook, Instagram and others. Social media allows users to share thoughts and ideas, as well as discuss current issues every day through various types of media such as text, photo, video, audio. Data from social media contains information related to activities carried out by the community and information when witnessing certain events [4]–[6]. Sentiment analysis is a popular method to find out public opinion on a problem or object. Many sentiment analysis studies have been carried out to monitor trends related to the spread of the Corona virus and its social impact in society [7]–[9].

This research developed a sentiment analysis model using TF-IDF and lexical method to identify public reactions due to COVID19. This paper consists of five chapters, including chapter one discussing the background, chapter two discussing the related research, chapter three discussing Sentiment Analysis method for public reaction identification, chapter four discussing the analysis

[1] *Department of Electrical Engineering, Faculty of Engineering, Semarang State University – Indonesia*
[2] *Department of Electrical Engineering, Faculty of Engineering, Semarang State University – Indonesia*
[3] *Department of Electrical Engineering, Faculty of Engineering, Semarang State University – Indonesia*
[4] *Department of Information System, Faculty of Computer Science, Universitas Dian Nuswantoro – Indonesia*
* *Corresponding Author Email: nur.iksan@mail.unnes.ac.id*

of experimental results, chapter five presenting conclusions.

## 2. Related Works

Several studies have succeeded in developing algorithms to analyze twitter sentiment from the COVID-19 problem. Study conducted by Machuca et al., (2021) developed machine learning algorithms and NLP techniques to infer whether public sentiment is positive or negative. A universal approach was used in this research, and it was possible to change the data source so that it could be analyzed to obtain the desired algorithm for processing text data from Twitter. The results obtained from this study showed that the algorithm developed has succeeded in classifying the types of sentiments. Machine learning algorithms and NLP techniques showed that 54% of users showed positive sentiments and 46% of users showed negative sentiments against COVID19 problem.

Alabid & Katheeth, (2021) developed two machine learning-based classification algorithms, Naive Bayes and Support Vector Machines, then pre-processing methods were applied in the algorithms to filter out unstructured tweets. The second algorithm was tested by classifying sentiments and opinions of individuals related to COVID19 vaccine on Twitter. Identification carried out in this study is expected to help public health institutions to find public opinion and direct vaccination promotion efforts. The tweet data processed in this study were 15,000 tweets. This study proves that the performance of the Naive Bayes algorithm outperforms the support vector machine algorithm.

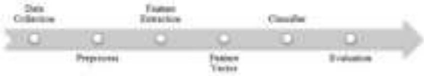
Deep Belief Neural Network (DBN) algorithm was implemented with pseudo labeling to classify tweets with Srikanth et al., (2022). This study also used a combination of preprocessing approaches such as tokenization, filtering, stemming, and building an N-gram model. The approach proposed in this study is expected to assist medical professionals and decision makers in determining the best course of action for each site based on their views on the pandemic. The results of this study indicated that the combination of the Deep Belief Neural Network classifier with bigram in the N-gram model outperforms other models. The selection of the correct features is proven to increase the accuracy of the algorithm.

Twitter sentiment analysis study on the COVID-19 case was analyzed by Efrilianda et al., (2021) using the fuzzy logic method. The fuzzy logic method has been successful in designing, creating and building bots that can analyze user opinions on Twitter. This study used 212 tweets for database. This data collection used random sampling method. The results showed that sentiment analysis during the COVID-19 period was still dominated by positive tweets. As many as 48% of positive tweets, 30% negative tweets, and 22% neutral tweets.

Hendrawati & Yanti, (2021) developed a Multi-layer Perceptron (MLP) algorithm using Backprogragation with Adam's optimization to analyze tweets related to COVID-19 in negative or neutral sentiments. There were 200 tweet databases used in this study. The results of this study indicated that the Multi-layer Perceptron (MLP) algorithm using Backprogragation with Adam's optimization developed has an accuracy of 70%, so this algorithm is considered successful in classifying sentiments for Indonesian-language tweets. Indonesian. This study analyzed sentiment and opinions related to the COVID-19 outbreak and then categorized them into positive.

This study succeeded in developing a sentiment analysis model to identify public reactions due to COVID-19. This study used Twitter social media to obtain data related to COVID-19 in Indonesia used as further identification based on the Sentiment Analytic approach. The data source was obtained from Twitter via the Twitter API. The features used were word choice, emoticons and polarity. In addition, Time of day and location were also used to map time and location. The tweet sentiment analysis method in this study has been developed using a data mining approach as a sentiment analysis model to identify public reactions to the COVID-19 issue.

## 3. Method



Sentiment analysis or opinion mining is an automatic textual data processing process that aims to obtain sentiment information contained in an opinion sentence. Sentiment analysis is done to sort out opinions or tendencies of a positive or negative opinion on a problem or object [9], [15]. A series of processes in sentiment analysis is shown in Figure 1:

**Fig. 1.** General Stages of Twitter Data Sentiment Analysis

Data Collection is the first stage in Sentiment Analysis. At the data collection stage, data can be obtained from the internet through various means, such as: web scraping, social media, news channels, E-commerce websites, forums, weblogs, and others [16]. In this study, data were obtained from public tweets on the Twitter application via the Twitter API.

Data pre-processing is the stage to convert the data into a basic form to ease the process in the next stage. Data pre-processing is an important technique for processing data mining performance. It includes several preprocessing tools available in data mining [17]. The preprocessing stage is divided into two parts, namely general functionality for tasks related to the text contained in tweets and specific functionality for tasks related to tweet characteristics [18]. Figure 2 shows the data preprocessing technique for tweet data used in this study:


**Fig. 2.** Pre-processing stages

1. Removing URL, tweets posted as text was filtered out to remove URL, such as "http://".
2. Removing special symbol in tweets posted in text, some symbols found in the tweets was removed, such as hashtag (#), retweet (RT) and username (@).

3. Tokenization, tweets posted in texts were split into phrases called as token using tokenize() function. This function was obtained from RegexpTokenizer() in the NLTK library.
4. Case folding, phrases produced by tokenization process were then changed into lower case letters using lower() function.
5. Nonstandard word handling, some words detected to be not standard, such as typo, abbreviation error, slang word, and lengthen words is returned into the standard forms.
6. Stemming, some words that have affixes, such as prefixes, insertions, suffixes and confixes will be returned to the basic word form. At this stage, Sastrawi library was used for Indonesian-language tweets, while NLTK library was used for English tweets. In addition, several stemming algorithms can be used were: Porter Stemmer (English & Indonesian), Nazief-Adriani Stemming (Indonesia), Arifin-Setiono Stemming (Indonesia), Khoja (Arabic). For example, the following libraries are used by Sastrawi (Stemming Nazief-Adriani) in the Stemming process:

```
# Stemming
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
factory      = StemmerFactory()
stemmer      = factory.create_stemmer()
input_stemm  = str(hasil_token)
hasil_stemm  = stemmer.stem(input_stemm)
```

Stopword Removal, some common words that do not have a significant effect on the sentence of the tweet will be removed using Sastrawi and NLTK libraries.

The next stage is feature extraction, which is the very basic stage in Sentiment analysis. In this technique, the text has to be converted into a feature vector with the help of a data-driven approach. TF-IDF is a popular method that is widely used to evaluate the importance of a word in a sentence. [19]. In this study, Feature Extraction was used based on The Frequency Inverse Document Frequency term or commonly referred to as TF-IDF. When calculating frequency term, a characteristic/feature is determined based on the most frequent words appear. The feature represents a document and has a heavy weight. It is different from TF-IDF, the word appearance frequency in a document is compared to the word appearance in all documents. If a word frequently appears in many documents, so the word is not used as a characteristic/feature because the word is considered as a general word and cannot represent a document. To obtain TF-IDF weight, the following Formulas (1) and (2) can be used:

$$tf.idf_{t,d} = tf_{t,d} \times idf_t \tag{1}$$

$$idf_t = \log \frac{N}{df_t} \tag{2}$$

In addition, the Feature Extraction stage uses the semantic approach to lexicons. The lexical method is a semantic approach that uses word meanings for feature extraction. In this approach, the more complete the lexicon sentiment used, the more precise the results obtained [20], [21]. This lexicon sentiment contains a list of words with positive polarity such as "good", "good", "safe" and negative words such as "bad", "slow", "sloppy". In the Feature extraction using the English lexicon, the first step done is doing the translation process using the Google Translate API and followed by calculating the sentiment score from sentiwordnet. Sentiment score calculations from sentiwordnet use Formulas 3 and 4. Feature extraction using the Indonesian lexicon takes advantage of the frequency of appearance of opinion words in each sentence that has positive and negative sentiments. The calculation of the sentiment score on the Indonesian lexicon uses the Formulas 5 and 6.

$$Score = \sum \frac{(PosScore - NegScore)}{tot_{index}} \tag{3}$$

$$Sentence_{Score} = \sum Score \tag{4}$$

$$Score = \sum frek_{opini_{word}} \tag{5}$$

$$Sentence_{Score} = \sum_{i=0}^{n} \sum Score_i \tag{6}$$

Where $PosScore$ denotes positive score of word, $NegScore$ denotes negatve score of word. $tot_{index}$ denotes the number of indexes of the word separated by characters #. $Score$ denotes weight score of the word (term). $Sentence_{Score}$ denotes sentiment score of sentences. $frek_{opini_{word}}$ denotes the frequency of occurrence of words containing opinion or sentiment.

This study combined two statistical approaches using TF-IDF and a semantic approach using lexical resources for weighting processes in feature extraction. Figure 3 below shows the sentiment analysis algorithm which is the proposed approach.

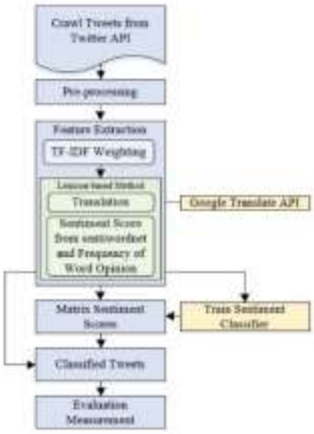



**Fig. 3.** Sentiment Analysis Algorithm Proposed

The classification process used the Naïve Bayes algorithm while the validation process was carried out by calculating the accuracy values. Prediction accuracy measurement was done by comparing the TFBS method to several algorithms used including ME, FBS, AFBS. Some of the methods used for comparison can be briefly described as follows:

1. ME: Maximum Entropy
2. FBS: Feature Based Sentiment using the Lexicon method
3. AFBS: Augmented Feature Based Sentiment using the Augmented Lexicon method
4. TFBS: TF-IDF and Lexicon Method

The accuracy value is obtained by comparing the amount of data from the classification results to the total amount 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 using formula 7 to 9. 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 to 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 reaction (happy, sad, angry, afraid, and love).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (7)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (8)$$

$$F - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (9)$$

## 4. Results and Discussion

This study used dataset collected from Twitter by inputting keyword "COVID19" using Indonesian language. The dataset was collected through crawling process using Tweepy libeary in Python with 3148 lines. The dataset went through data normalization. Duplicated data found in tweets that are retweeted (RT) by other users were also removed. After that, the dataset were labelled manually based on 5 reaction categories: happy, sad, angry, fear, and love. The following table presents the 5 categories for tweet text label.

**Table 1.** Labelling tweet text in the dataset

| Original Tweet | Tweet in English | Labels |
|---|---|---|
| Alhamdulillah ada rejeki aku dan papaku juga udah aware bakal ribet kalau mengandalkan full support pemerintah. Jadi skrg segala macem supply obat, vitamin bahkan konsul online beneran modal sendiri. Jadi beneran, covid is real gengs. Stay at home, stay safe. Gbu! | Thank God for all fortune, my dad and I realize that it will be more complicated if we only rely on the government's support. So now I supply my own drugs and vitamins and have my own online consultation. So really, covid is real guys. Stay at home, stay safe. Gbu! | Happy |
| tiap hari dengarin orang sesak nafas karna covid sedih banget dengarnya :( beruntung bgt gua ga separah mereka | every day I hear many news that people are suffering from shortness of breath caused by covid, I am so sad :( but I'm so lucky I do not suffer as bad as they do | Sad |
| @rinnelikers Iya ih,sampe ada yang gak percaya kalo covid itu ada,padahal tinggal ikutin peraturan.Pakai masker,jangan kumpul kumpul... :( | @rinnelikers absolutely, I wonder why there are people who dont believe that covid is real, they just have to follow the rules, wearing mask and maintain social distancing :( | Angry |
| kasus covid makin ngeri dalem waktu dekat pengen nekat ke purworejo tapi tetangga ada yg kena covid dan bbrp hari yg lalu siweran ketemu aku di warung | The covid case is getting scary, actually I desperatly want to go to Purworejo, but a few days ago, my neighbors are infected and I met them at the shop | Fear |
| Mari kita sama sama doakan agar pesakit covid-19 semua diberikan kekuatan dan kesabaran untuk melawan virus ini. Amin | Let's pray together that all of the Covid-19 patients will be given strength and patience to fight this virus. Amen | Love |

Tweet text contains irrelevant information or noise, such as hashtag (#), html (http://www), mention (@), number, the use of local language and abbreviation. The first stage of sentiment analysis is pre-processing to remove URL, special symbol, punctuation, tokenization, case folding, steming as presented in Figure 2. To remove irrelevant information, URL, special symbol of tweet, ASCII symbol, numbers and punctuation were removed. Figure 4 is an example of tweets still containing noise. Therefore, pre-processing is required in this tweet.


**Fig. 4.** Example of tweet in Twitter

Pre-processing was carried out in the tweet text presented in Figure. 4 and the result is presented in Table 2.

**Tabel 2.** Stage of Removing in Pre-processing Data

| Pre-processing stage | Original Tweet | Tweet in English |
|---|---|---|
| Tweet | "Kalo di kudus ada 300 nakes yang tertular covid19, menurut kalian yang salah dan lalai itu siapa? Nakes juga harus diingatkan terus jangan lalai, jangan hanya warganya aja. Sekarang kan APD sdh lengkap dan mudah jadi kagak ada alasan lagi kaya tahun lalu.." | "in Kudus, there are 300 health workers who have infected covid19, who do you think is wrong and negligent? Health workers must also be reminded not to be negligent, not just the citizens. Now, the PPE is complete and available, so there is no reason to do the same mistake like last year.." |
| Preprocessing Results | "kudus nakes tular covid salah lalai nakes lalai warga apd lengkap mudah kagak alasan kaya" | "kudus health worker infected covid wrong negligent health worker citizen apd complete available mo reason like " |
| Tokenization | [kudus] [nakes] [tular] [covid] [salah] [lalai] [nakes] [lalai] [warga] [apd] [lengkap] [mudah] [kagak] [alasan] [kaya] | [kudus] [health worker] [infected] [covid] [wrong] [negligent] [health worker] [citizen] [apd] [complete] |

| | | |
|---|---|---|
| | | [available] [no] [reason] [like] |

The features used were matrix score from TF-IDF and lexicon-based method. In the pre-processing process, several tasks include removing URL, converting emoticons, removing special symbol of tweets, removing ASCII symbol numbers and punctuation, tokenization, case folding, non-standard word handling, stemming, stop word removal.

The results of the accuracy test are shown in Figure 5. In these results, there are 5 categories of each method, namely happy, sad, fear, angry and love. Furthermore, we can see that, the proposed method is more accurate than the other methods.

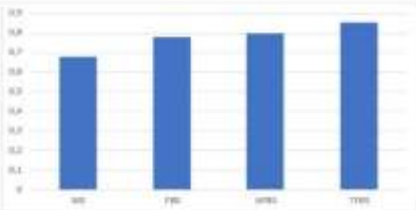


**Fig. 5.** Accuracy Results of Each Method

Furthermore, the results of the Recall, Precision and F-Score measurements of each method can be seen in Figure 6. In these results, the ME method has poor performance while AFBS and FBS have better performance with superior AFBS. Our proposed TFBS method outperforms AFBS. So, TFBS has better performance than other methods because it can correctly identify and classify tweets with opinions.

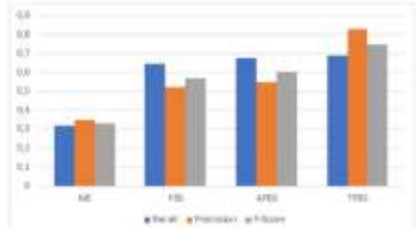
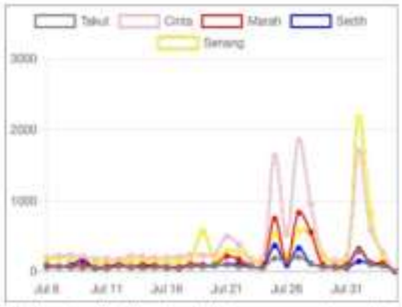


**Fig. 6.** Result of the Recall, Precision and F-Score measurements of each method

The next experiment was to classify the trend of public reactions in facing the COVID19 pandemic using the TFBS method. Classifications of public reactions are happy, sad, fear, angry and love. Based on the data from the crawling results obtained, the public reaction is presented as shown in Figure 7. Public Reaction of "love" was higher than other reactions. This was because in 24 Juli- 31 Juli 2022 the people's reaction to the COVID19 pandemic in Indonesia still seemed to be normal and they tended to think that this virus cannot attack Indonesia because of the on-going vaccination program and the creation of herd immunity. They though that covid virus only attacts those who do not follow health protocol and do not co vaccination. Besides, trend of public reaction for "angry" was in the rank second after "love" on July 25-27 2022. This was because in some areas there was an increase in Covid patients because of community activities that have returned to normal, including schools that have implemented




offline learning. Besides, there is also issue on new virus arising in the community which is monkeypox.

**Fig. 7.** Trend of public reactions from 06/07/2022 to 04/08/2022

In addition, trend of public reaction was also visualized in real time on the dashboard of public reaction monitoring system with a time period of the last 1 hour. In the visualization, we observed the trend of public reaction every 2 minutes. For instance, in Figure 8, the trend of public reaction from 05.20 AM to around 06.00 showed a fear reaction. Therefore, based on the trend we could conclude opinions related COVID19 that was being discussed within the


society.

**Fig. 8.** Trend of public reactions in real time in the last 1 hour

After that, qualification results were shown in a visualization of word cloud as presented in Figure 9 to Figure 13. The visualization shows the words that appear frequently in each category of public reactions regarding covid19. In each of these categories, the words that came up frequently in the "happy" category were: health, protocol and discipline; "sad": infected, sick, fever; "love": health,

protocol, discipline; "fear": results, examination, ministry of health; and "angry": infected, vaccinated, positive.

Fig. 9. Word Cloud of popular topic of public reactions for the sad category

In Figure 9, the trend of public reactions for the happy category in the word cloud consists of several words that frequently appeared and became opinions widely discussed in the community. The words were: Protocol, Health, Masks and Discipline. These words became a trend in the community because at that time, the enforcement of health protocol discipline was still the priority and had become public awareness.


Fig. 10. Word cloud of popular topic of public reactions for the sad category

In Figure 10, the trend of public reactions for the sad category was dominated by words such as: infected, sick, fever, positive and cough. These words became a trend in society because at that time people started working from office and schools were carried out offline. Therefore, many of COVID 19 cases and other symptoms that lead to COVID19 were reported.


Fig. 11. Word cloud of popular topic of public reactions for the love category

In Figure 11, some words that appeared such as: Health, protocol, discipline, progress, increase, deltacron, variant and mask. These words became a trend in society with the category of love. The trend in this category is almost the same as the Happy category. However, this trend in the love category shows an opinion which means hope for the better. During this period, new variants of COVID-19 began to appear, such as omicron and deltacron.


However, the public expressed their love reaction while maintaining health protocols and discipline in wearing masks.

Fig. 12. Word Cloud of Popular topic of public reactions for the fear category

In Figure 12, the trend of public reactions for the fear category is dominated by words such as: ministry of health, health, laboratory, examination, discipline and protocol. During this period, public opinion was more focused on the issue of an increase in COVID-19 patients, the emergence of new COVID variants and the emergence of the monkeypox virus so that a policy from the Ministry of Health was needed to respond this case. These issues are the reaction of the public to always be vigilant and afraid so that they remind each other to always be disciplined in health protocols.

In Figure 13, the trend of public reactions for the angry category is dominated by words such as: sick, infected, positive, death and vaccine. The cause of this reaction is almost the same as the sad category.


Fig. 13. Word Cloud of Popular topic of public reactions for the angry category

Table 3. Word Cloud of Popular Topic from Indonesian User

| Label | Original World Cloud | English of World Cloud |
|---|---|---|
| Happy | "vaksinasi, pemalsuan, adhanom, masker angka, kesehtan, test, sosial, ulang, tedros, direktur, protokol, masuk, desak, disiplin, hasil, perkuatjagaimunitas, record menerapkan" | " vaccination, forgery, adhanom, figure mask, health, test, social, repeat, tedros, director, protocol, admission, urge, discipline, results, strengthen immunity, record, apply " |
| Sad | "sekolah, batuk, sakit, orang, udah, bener, makan, demam, sedih, anak konser, positif" | " school, cough, sick, people, already, right, eat, infected, fever, sad, concert, children, positive " |
| love | "Omicron, kenaikan, disiplin, protokol, masker, perkuatjagaimunitas, kesehatan, varian, naik, deltacron, prokes, subvarian" | "Omicron, increase, discipline, protocol, mask, strengthen immunity, health, variant, up, deltacron, prokes (health protocol), subvariant" |
| Fear | "Laboratorium, kemenkes, kesehatan, pemeriksaan, sakit, masuk, hasil, kena, disiplin, protokol" | "Laboratory, Ministry of Health, health, examination, sick, admission, result, infected, discipline, protocol" |
| Angry | "Sakit, banget, virus, kena, positif, kematian, vaksin, jumlah" | "Sick, desperately, viral, infected, positive, death, vaccine, number" |

Table 3 shows some words in the wordcloud from each category. These words appear because they have a high frequency of occurrence of tweet texts posted by users. So, opinions that develop in the community can be identified.

Furthermore, Figure 14 presents the visualization of the report from the monitoring system of public reaction to COVID19. This figure informs the tweet recapitulation from each reaction category. In addition, it is also able to show a graph of people's reactions in real time or at a certain time that has been determined before.

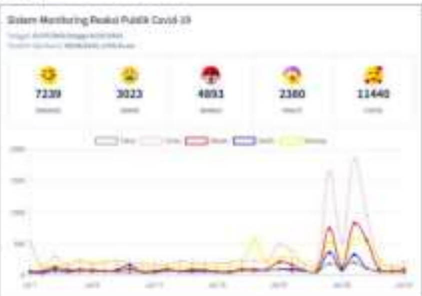



**Fig. 14.** Visualization of the report from the monitoring system of public reaction to COVID19

## 5. Conclusion

This study has developed a sentiment analysis model using TF-IDF and lexical methods to identify public reactions due to COVID19. Prediction accuracy measurement result shows our proposed method is more accurate than the other methods which are 0,854. Furthermore, the results of the Recall, Precision and F-Score measurements shows the ME method has poor performance while AFBS and FBS have better performance with superior AFBS. Our proposed TFBS method outperforms AFBS. TFBS has better performance than other methods because it can correctly identify and classify tweets with opinions. Experiment was conducted to classify the trend of public reactions in dealing with the COVID19 pandemic using the TFBS method, some of category for public reaction was identified and visualized on the dashboard of monitoring system of public reaction to COVID19.

## Acknowledgements

This study was supported/partially supported by The Directorate of Research and Community Service – Ministry of Research, Technology and Higher Education (DRPM – Kemenristekdikti); and Semarang State University (UNNES). Thanks for their assistance in facilitating research.

## References

[1] R. P. Pratama and A. Tjahyanto, "The influence of fake accounts on sentiment analysis related to COVID-19 in Indonesia," in *Procedia Computer Science*, 2021, vol. 197, no. 2021, pp. 143–150. doi: 10.1016/j.procs.2021.12.128.

[2] A. D. Fakhira, A. S. Pawitra, K. C. Diyanah, E. A. Wikurendra, I. Nagy, and N. S. M. Abdeljawad, "Awareness of Doing 3M (Wearing Mask, Physical Distancing, Washing Hands) During Pandemic Era in Rural and Urban Families," *Jurnal Kesehatan Lingkungan*, vol. 13, no. 2, p. 94, 2021, doi: 10.20473/jkl.v13i2.2021.94-101.

[3] A. K. and M.-V. Dorothée Allain-Dupré, Isabelle Chatry and Michalun, "The territorial impact of COVID-19: Managing The Crisis Across Levels of Government," 2020.

[4] L. Nahar, Z. Sultana, N. Iqbal, and A. Chowdhury, "Sentiment Analysis and Emotion Extraction: A Review of Research Paradigm," *1st International Conference on Advances in Science, Engineering and Robotics Technology 2019, ICASERT 2019*, no. December, 2019, doi: 10.1109/ICASERT.2019.8934654.

[5] Z. Drus and H. Khalid, "Sentiment analysis in social media and its application: Systematic literature review," *Procedia Computer Science*, vol. 161, pp. 707–714, 2019, doi: 10.1016/j.procs.2019.11.174.

[6] A. Pertiwi, A. Triayudi, and E. T. E. Handayani, "Sentiment Analysis of the Impact of Covid-19 on Indonesia's Economy through Social Media Using the ANN Method," *Jurnal Mantik*, vol. 4, no. 1, pp. 605–612, 2020.

[7] P. Yadav and D. Pandya, "SentiReview: Sentiment Analysis based on Text and Emoticons," in *International Conference on Innovative Mechanisms for Industry Applications*, 2017, pp. 467–472.

[8] S. Peng, G. Wang, and D. Xie, "Social Influence Analysis in Social Networking Big Data: Opportunities and Challenges," *IEEE Network*, vol. 31, no. 1, pp. 11–17, 2017, doi: 10.1109/MNET.2016.1500104NM.

[9] A. Go, L. Huang, and R. Bhayani, "Twitter Sentiment Analysis," 2009.

[10] C. R. Machuca, C. Gallardo, and R. M. Toasa, "Twitter sentiment analysis on coronavirus: Machine learning approach," *Journal of Physics: Conference Series*, vol. 1828, no. 1, 2021, doi: 10.1088/1742-6596/1828/1/012104.

[11] N. N. Alabid and Z. D. Katheeth, "Sentiment analysis of twitter posts related to the covid-19 vaccines," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 24, no. 3, pp. 1727–1734, 2021, doi: 10.11591/ijeecs.v24.i3.pp1727-1734.

[12] J. Srikanth, A. Damodaram, Y. Teekaraman, R. Kuppusamy, and A. R. Thelkar, "Sentiment Analysis on COVID-19 Twitter Data Streams Using Deep Belief Neural Networks," *Computational Intelligence and Neuroscience*, 2022, doi: https://doi.org/10.1155/2022/8898100.

[13] D. A. Efrilianda, E. N. Dianti, and O. G. Khoirunnisa, "Analysis of twitter sentiment in COVID-19 era using fuzzy logic method," *Journal of Soft Computing Exploration*, vol. 1, no. 1, pp. 1–5, 2021, doi: 10.52465/joscex.v2i1.12.

[14] T. Hendrawati and C. P. Yanti, "Analysis of Twitter Users Sentiment against the Covid-19 Outbreak Using the Backpropagation Method with Adam Optimization," *Journal of Electrical, Electronics and Informatics*, vol. 5, no. 1, p. 1, 2021, doi: 10.24843/jeei.2021.v05.i01.p01.

[15] T. N. Simanjuntak and S. Pramana, "Sentiment Analysis on Overseas Tweets on the Impact of COVID-19 in Indonesia," *Indonesian Journal of Statistics and Its Applications*, vol. 5, no. 2, pp. 304–313, 2021, doi: 10.29244/ijsa.v5i2p304-313.

[16] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, 2022, doi: https://doi.org/10.1007/s10462-022-10144-1.

[17] T. N. Prakash and A. Aloysius, "Data Preprocessing in Sentiment Analysis using Twitter Data," *International Education Applied Research Journal*, vol. 3, no. 7, 2019.

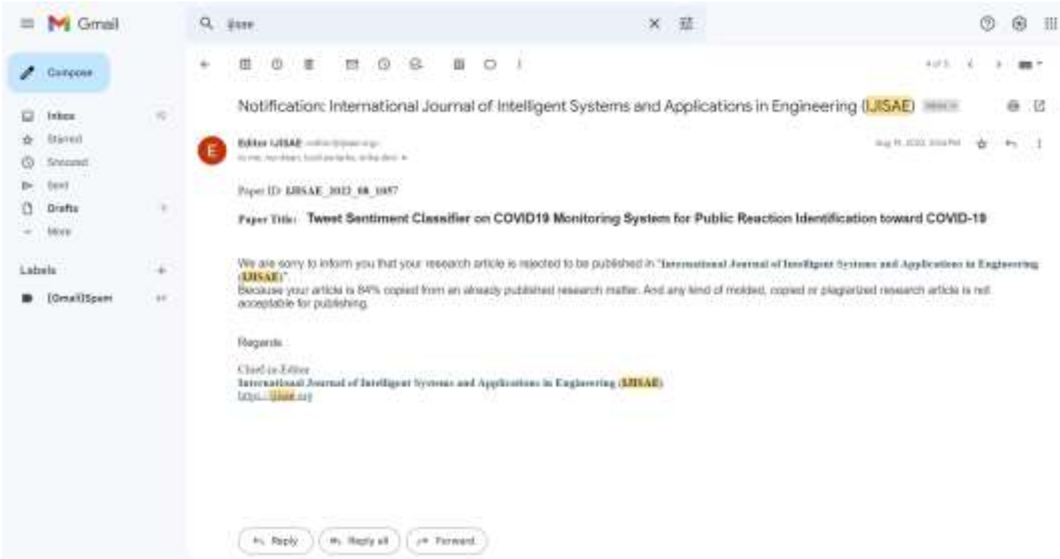
[18] A. F. Hidayatullah and M. R. Ma'arif, "Pre-processing Tasks in Indonesian Twitter Messages," *Journal of Physics: Conference Series, Volume 801, International Conference on Computing and Applied Informatics - IOP Publishing Ltd*, 2016, doi: 10.1088/1742-6596/801/1/012072.

[19] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," *Procedia Computer Science*, vol. 152, pp. 341–348, 2019, doi: 10.1016/j.procs.2019.05.008.

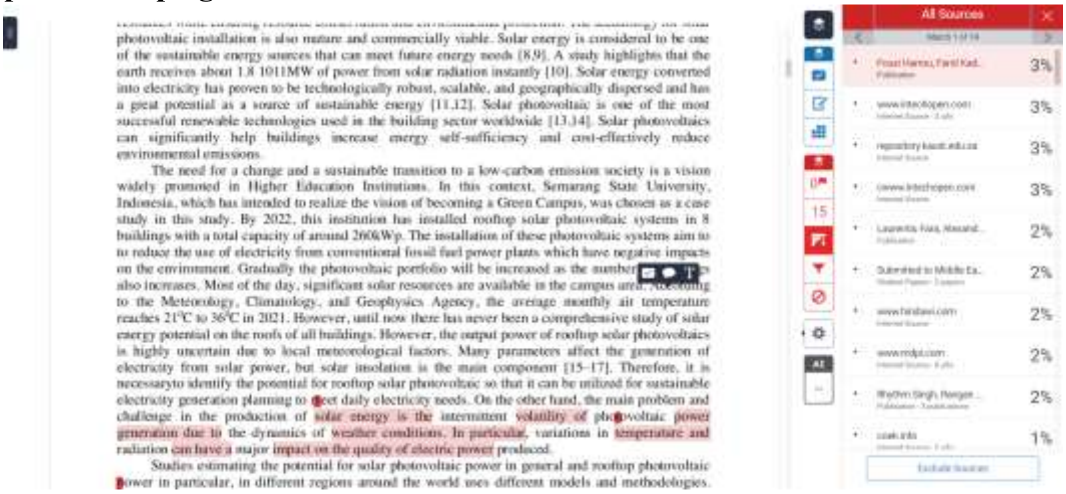
[20] Kusrini and M. Mashuri, "Sentiment Analysis In Twitter Using Lexicon Based and Polarity Multiplication," 2019, doi: 10.1109/ICAIIT.2019.8834477.

[21] N. Medagoda, S. Shanmuganathan, and J. L. Whalley, "Sentiment Lexicon Construction," 2015.

5. **Pemberitahuan via email bahwa naskah ditolak dengan alasan persentase plagiarism Turnitin melebihi standart IJISAE.**



6. **Proof-readiang naskah dan pengecekan Turnitin, revisi untuk penurunan persentase plagiarism.**


photovoltaic installation is also mature and commercially viable. Solar energy is considered to be one of the sustainable energy sources that can meet future energy needs [8,9]. A study highlights that the earth receives about 1.8 1011MW of power from solar radiation instantly [10]. Solar energy converted into electricity has proven to be technologically robust, scalable, and geographically dispersed and has a great potential as a source of sustainable energy [11,12]. Solar photovoltaic is one of the most successful renewable technologies used in the building sector worldwide [13,14]. Solar photovoltaics can significantly help buildings increase energy self-sufficiency and cost-effectively reduce environmental emissions.

The need for a change and a sustainable transition to a low-carbon emission society is a vision widely promoted in Higher Education Institutions. In this context, Semarang State University, Indonesia, which has intended to realize the vision of becoming a Green Campus, was chosen as a case study in this study. By 2022, this institution has installed rooftop solar photovoltaic systems in 8 buildings with a total capacity of around 260kWp. The installation of these photovoltaic systems aim to to reduce the use of electricity from conventional fossil fuel power plants which have negative impacts on the environment. Gradually the photovoltaic portfolio will be increased as the number [illegible] also increases. Most of the day, significant solar resources are available in the campus area. According to the Meteorology, Climatology, and Geophysics Agency, the average monthly air temperature reaches 21°C to 36°C in 2021. However, until now there has never been a comprehensive study of solar energy potential on the roofs of all buildings. However, the output power of rooftop solar photovoltaics is highly uncertain due to local meteorological factors. Many parameters affect the generation of electricity from solar power, but solar insolation is the main component [15–17]. Therefore, it is necessaryto identify the potential for rooftop solar photovoltaic so that it can be utilized for sustainable electricity generation planning to meet daily electricity needs. On the other hand, the main problem and challenge in the production of solar energy is the intermittent volatility of photovoltaic power generation due to the dynamics of weather conditions. In particular, variations in temperature and radiation can have a major impact on the quality of electric power produced.

Studies estimating the potential for solar photovoltaic power in general and rooftop photovoltaic power in particular, in different regions around the world uses different models and methodologies.

International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

# Tweet Sentiment Classifier on COVID-19 Monitoring System for Public Reaction Identification using Hybrid Feature Extraction Method

**Djoko Adi Widodo[1], Nur Iksan*[2], Budi Sunarko[3], Erika Devi Udayanti[4]**

Submitted: xx/xx/202x Accepted : xx/xx/202x DOI: [illegible]

***Abstract:*** Coronavirus (COVID-19) outbreak has currently become a public concern. The virus has been declared by WHO as a pandemic due to its spread to all countries, including Indonesia. Several prevention strategies have been implemented such as social distancing, quarantine, working from home, using masks, washing hands, and others. However, these steps have not been effective because not all people are aware and disciplined to apply in daily life and this means there is a potential threat for the spread of the virus. It was also observed that the pandemic greatly affected other areas besides the health sector ranging from the social, political, religious, and economic aspects to the resilience of the people. These can be observed through direct observation of the community or activities of the people on social media, especially in relation to the socio-economic aspect. Therefore, this research was conducted using social media, specifically Twitter, via the Twitter API to obtain data related to the COVID-19 pandemic in Indonesia. The data were mined and analysed to identify trends related to the spread of the Coronavirus and its social impacts on society through public reactions. Moreover, a sentiment analysis model was developed using TF-IDF and lexical methods based on matrix scores features. The prediction accuracy measurement showed that the proposed method is more accurate than the others as indicated by 0.854. Furthermore, the Recall, Precision, and F-Score measurements showed that the ME method has poor performance while AFBS and FBS have better performance.

***Keywords:*** *Tweeter Analysis, Sentiment Classifier, TF-IDF, Lexicon Based, COVID-19*

## 1. Introduction

World Health Organization (WHO) declared the COVID-19 virus a pandemic due to its spread to all countries in the world, thereby, becoming a serious concern for the public, government, and world health institutions [1]. The virus started emerging in Indonesia in early January 2020 and the data from the handling team showed an increasing trend with nearly 600,000 people reported to be exposed and approximately 18,000 died.

Several strategies were implemented to prevent further spread such as washing hands, wearing masks, quarantine, social distancing, working from home, and others [2]. However, these steps were not effectively implemented in the community due to low public awareness and lack of discipline in daily life and this indicated a potential threat of continuous exposure to the virus. This led the government to introduce another strategy which involved launching a vaccination program for all elements of society to prevent infection.

This vaccination program was one of the effective efforts made to suppress the spread with the main goal of providing immunity against COVID-19 virus infection and ensuring the formation of herd or group immunity. Furthermore, the spread was also prevented by providing information to increase public awareness on the best way to conduct their activities during the period.

In addition to the health sector, the pandemic also significantly affected the economic, religious, security, resilience, political, and social sectors [3]. These were not identified directly as the case with the health sector but through the observation of community activities using different media such as Instagram, Facebook, Twitter, and others. These media serve as social sensors for communities through the uploading of ideas, thoughts, and opinions in the forms of text, video, photos, and audio. The contents are focused on activities and conditions of the things the people feel and see in their surrounding environment which are subsequently sha red on social media platforms. Meanwhile, sentiment analysis can be used to identify and predict future conditions using the information [4-6].

A sentiment analysis method was developed in this research to identify public opinion related to the spread of the COVID-19 virus and its social impact on society [7–9]. This was achieved using the TF-IDF and Lexical methods for feature extraction and Naive Bayes for classification.

## 2. Related Works

Sentiment analysis has become an interesting topic and its methods have been widely developed, especially on Twitter concerning the COVID-19 spread. It was used by Machuca [10] to determine negative and positive sentiments through Machine Learning (ML) and Natural Language Processing (NLP) methods. The Logistic Regression Algorithm was used as an ML algorithm to classify feelings or emotions regarding the COVID-19 topic on Twitter in 2020. Meanwhile, the TF-IDF method was used to extract NLP features and the classification accuracy value was found to be 78.5%. Alabid & Katheeth, (2021) also developed a model

[1] *Department of Electrical Engineering, Faculty of Engineering, Semarang State University – Indonesia*
[2] *Department of Electrical Engineering, Faculty of Engineering, Semarang State University – Indonesia*
[3] *Department of Electrical Engineering, Faculty of Engineering, Semarang State University – Indonesia*
[4] *Department of Information System, Faculty of Computer Science, Universitas Dian Nuswantoro – Indonesia*
* *Corresponding Author Email: nur.iksan@mail.unnes.ac.id*

framework to classify public sentiment and opinion regarding the implementation of the COVID-19 vaccination program. The research further compared 2 ML methods including Support Vector Machine (SVM) and Naive Bayes (NB) to classify sentiment polarity into positive, negative, and neutral. The experimental results showed that the NB method had a higher accuracy performance than SVM with 0.81 and 0.75 respectively. Srikanth et al., (2022) also conducted a sentiment analysis using a combination of preprocessing methods and word embedding to perform feature extraction. The Deep Belief Neural Network (DBN) algorithm was used to classify tweets through a pseudo-labeling process and the experimental results showed that the DBN algorithm achieved an accuracy value of 90.3% which is better than the other methods. Moreover, Hendrawati & Yanti, (2021) developed a sentiment analysis method to determine positive, negative, and neutral polarity using the Backpropagation Multi-Layer Perceptron (MLP) algorithm and the results showed an accuracy value of 70%.

It was observed that these studies did not consider the development of a feature extraction method as an important factor in increasing the accuracy of sentiment classification. They also only classified sentiments into negative, positive, and neutral polarities instead of emotions and feelings. Therefore, this research developed a feature extraction method by combining TF-IDF with the lexical method. Sentiment analysis was also developed concerning the community's reactions to the COVID-19 pandemic with a focus on happy, sad, angry, fear, and love categories.

## 3. Method

Sentiment analysis is also known as opinion mining and it is normally used to process natural discussion or text commonly referred to as NLP to determine the sentiments embedded in an opinion. It usually categorizes the opinion sentences in the form of polarity which include Negative, Positive, and Neutral [9], [15]. It can also be used for classification in other forms such as happy, sad, angry, fear, and love. The steps in the sentiment analysis process are shown in the following Figure 1:

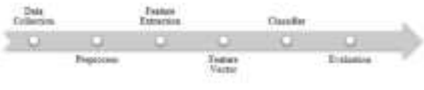



**Fig. 1.** General Stages of Twitter Data Sentiment Analysis

The sentiment analysis consists of 6 stages of data collection, data pre-processing, feature extraction, vectorization, classification, and evaluation. The data collection stage was used to retrieve text data from Twitter through the API [16]. The data pre-processing was used to prepare the collected data for extraction into features [17] and this is important to determine the performance of the sentiment analysis process later. Overall, this stage is divided into two which include general pre-processing associated with the text contained in the tweet and special pre-processing related to the characteristics of the tweet text [18]. Figure 2 shows the stages of data pre-processing conducted in this research.


**Fig. 2.** Pre-processing stages

1. Removal URL by filtering the tweet text to remove URLs such as "HTTP://".
2. Removal of the special symbol in the tweets such as hashtags (#), retweets (RT), and usernames (@).
3. Tokenization such that the tweet text in the form of a sentence was be divided into several constituent words or phrases called tokens using the tokenize() function. The function was obtained from the RegexpTokenizer() package in the NLTK library.
4. Case folding which involved changing some words from the tokenization process to lowercase letters using the lower() function.
5. Nonstandard word handling which involved returning the detected unstandardized words such as errors in the use of spelling, abbreviations, slang words, and elongated words to standard form.
6. Stemming which involved returning some words with affixes such as prefixes, insertions, suffixes, and confixes to their basic word form. At this stage, the literary library was used for Indonesian-language tweets while the NLTK library was used for English tweet text. Furthermore, several stemming algorithms were also applied such as Porter Stemmer (English & Indonesian), Nazief-Adriani Stemming (Indonesia), Arifin-Setiono Stemming (Indonesia), and Khoja (Arabic). For example, the following libraries were used by Sastrawi (Stemming Nazief-Adriani) in the Stemming process:

```
# Stemming
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
factory      = StemmerFactory()
stemmer      = factory.create_stemmer()
input_stemm  = str(result_token)
result_stemm = stemmer.stem(input_stemm)
```

7. Stopword Removal stage which involved removing common words that do not have a significant effect on the tweet sentence. This was achieved through the Literature and NLTK libraries.

The next stage is extraction which is very fundamental in the sentiment analysis process [19]. It involves converting the pre-processed text into vector form and was conducted in this research using the TF-IDF approach which normally extracts sentence features based on the frequency of word occurrence. Moreover, the frequency term was calculated by determining a feature based on the consistent appearance of a word. The features represent a document and have a high weight. In contrast to TF-IDF, the frequency of a word occurrence in a document is usually compared with an entire database. In a situation the word appears too often in several documents, it cannot be used as a feature because it is considered a general word that does not represent a specific document. The TF-IDF weight was calculated using the following formula:

$$tf.idf_{t,d} = tf_{td} \times idf_t \tag{1}$$

$$idf_t = \log \frac{N}{df_t} \tag{2}$$

The lexical method was also applied in this research and it involves using the word meaning to determine the sentence features. The accuracy of the sentiment analysis is higher when the word meaning is more complete [20], [21]. The use of this method usually focuses on applying word lists such as "clean", "healthy", "good", and "safe" to determine positive polarity and others such as "sloppy", "bad", "and slow" to determine the negative polarity. This feature extraction for the English lexicon was conducted using the library from the Google Translate API and the lexicon was translated first after which the sentiment scores were calculated from sentiwordnet using Equations (3) and (4). Meanwhile, the feature extraction of the Indonesian language lexicon was conducted by calculating the appearance of the word opinion in each sentence after which the positive and negative sentiments were determined using Equations (5) and (6).

$$Score = \sum \frac{(PosScore - NegScore)}{tot_{index}} \tag{3}$$

$$Sentence_{Score} = \sum Score \tag{4}$$

$$Score = \sum frek_{opini_{word}} \tag{5}$$

$$Sentence_{Score} = \sum_{i=0}^{n} \sum Score_i \tag{6}$$

This means TF-IDF and Lexical methods which are statistical and semantic approaches respectively were combined in this research to determine the weighting of the feature extraction process. The proposed sentiment analysis model is presented in the following Figure 3.

At the classification stage, the Naive Bayes algorithm was used to determine the tweets consisting of 5 classes which include fear, angry, love, sad, and happy. Furthermore, the performance evaluation was conducted by comparing the Feature-Based Sentiment algorithm using TF-IDF and Lexicon (TFBS) with several other methods such as Maximum Entropy (ME), Feature-Based Sentiment using the lexicon (FBS) method, and Augmented Feature-Based Sentiment using the augmented lexicon (AFBS) method.

Performance evaluation was conducted by calculating the accuracy, precision, recall, and F-Score. The accuracy value was calculated by comparing the number of correct data and the total amount of data. A higher accuracy value indicates a better method. Moreover, precision, recall, and F-score were measured to determine any deviations in the data. Precision value was determined by comparing the number of relevant classification data with the total amount of data from a particular class. The recall value was obtained from the comparison of the amount of relevant data with the total relevant data while the F-Score value was determined by calculating the average value of precision and recall. The equations used to calculate precision, recall, and f-score values are stated as follows.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{7}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{8}$$

$$F - Score = 2 * \frac{Precision * Recall}{Precission + Recall} \tag{9}$$

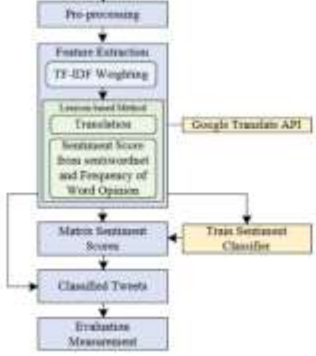



**Fig. 3.** Proposed Sentiment Analysis Model

## 4. Results and Discussion

This research used a dataset obtained through Twitter using the keyword "COVID-19" in Indonesian. The dataset was retrieved through a crawling process using the Tweepy library in Python for as many as 3148 lines. It was subsequently passed through the normalization stage to eliminate duplicated data such as the tweets reposted by other users or retweets (RT). The dataset was manually labelled using 5 categories of reactions which include happy, sad, angry, fear, and love. The following table provides examples of 5 categories in tweet text labelling.

**Table 1.** Labelling of tweet text in the dataset

| Original Tweet | Tweet in English | Labels |
|---|---|---|
| Alhamdulillah ada rejeki aku dan papaku juga udah aware bakal ribet kalau mengandalkan full support pemerintah. Jadi skrg segala macem supply obat, vitamin | Thank God, I have money and my father is also aware that it will be difficult to rely on full support from the government. Therefore, | Happy |

| | | |
|---|---|---|
| bahkan konsul online beneran modal sendiri. Jadi beneran. covid is real gengs. Stay at home, stay safe. Gbu! | now all kinds of supplies of drugs, vitamins, and even online consuls are real capital. Covid is real. Stay at home, stay safe. Gbu! | |
| tiap hari dengarin orang sesak nafas karna covid sedih banget dengarnya :( beruntung bgt gua ga separah mereka | Every day, I see people with breath shortness because of Covid, it's really sad to hear :( I'm so lucky, I'm not as severe as them | Sad |
| @rinnelikers Iya ih,sampe ada yang gak percaya kalo covid itu ada,padahal tinggal ikutin peraturan,Pakai masker,jangan kumpul kumpul... :( | @rinnelikers Yes, some people don't believe that covid exists, even though all you have to do is follow the rules, wear a mask, avoid crowdedness... :( | Angry |
| kasus covid makin ngeri dalam waktu dekat pengen nekat ke purworejo tapi tetangga ada yg kena covid dan bbrp hari yg lalu sliweran ketemu aku di warung | The Covid case is getting scary. In the near future, I want to go to Purworejo, but a neighbour has been infected and a few days ago he met me at the shop | Fear |
| Mari kita sama sama doakan agar pesakit covid-19 semua diberikan kekuatan dan kesabaran untuk melawan virus ini. Amin | Let's all pray that all Covid-19 patients are given strength and patience to fight this virus. Amen | Love |

The tweets contain some useless information or noise such as hashtags (#), Html (http://www), mentions (@), numbers, local language usage, and abbreviations. The initial stage in the sentiment analysis process is the pre-processing which involves removing the URL, special symbol, punctuation, tokenization, case folding, and stemming as shown in Figure 2. The useless information was removed by omitting URLs, special symbols, ASCII numbers, and punctuations. Figure 4 shows an example of a tweet containing noise and this means there is a need for pre-processing.


**Fig. 4.** Example of a tweet on Twitter

The tweet text in Figure 4 was pre-processed and the results are presented in the following Table 2.

**Table 2.** Cleansing at the Pre-processing Stage

| Pre-processing stage | Original Tweet | Tweet in English |
|---|---|---|
| Tweet | "Kalo di kudus ada 300 nakes yang tertular covid19, menurut kalian yang salah dan lalai itu siapa? Nakes juga harus diingatkan terus jangan lalai, jangan hanya warganya aja. Sekarang kan APD sdh lengkap dan mudah jadi kagak ada alasan lagi kaya tahun lalu.." | "In Kudus, there are 300 health workers who have contracted Covid19, who is wrong and negligent? Health workers should also be reminded not to be negligent, not just citizens. Now, the PPE is complete and easy, hence there's no excuse like last year." |
| Pre-processing Results | "kudus nakes tular covid salah lalai nakes lalai warga apd lengkap mudah kagak alasan kaya" | "Kudus health workers infected covid wrong negligent health workers negligent citizens PPE complete easy no excuse like" |
| Tokenization | [kudus] [nakes] [tular] [covid] [salah] [lalai] [nakes] [lalai] [warga] [apd] [lengkap] [mudah] [kagak] [alasan] [kaya] | [Kudus] [health workers] [infected] [covid] [wrong] [negligent] [health workers] [negligent] [citizens] [PPE] [complete] [easy] [no] [excuse] [like] |

Several processes were implemented at the pre-processing stage and these include the deletion of emoticons, URLs, special symbols, ASCII symbols, and punctuation marks. Furthermore, the text characteristics were handled using different methods such as tokenization, case folding, non-standard word handling, stemming, and stop-word removal. The results were later forwarded to the feature extraction stage to produce a scoring matrix by combining the TF-IDF and lexicon methods. Figure 5 shows the accuracy values obtained from using different feature extraction methods. It is important to note that the focus was placed on the tweet classification using five classes of fear, angry, love, sad, and happy. The results showed that the proposed TFBS method had a higher accuracy value of 0.85 compared to other methods.

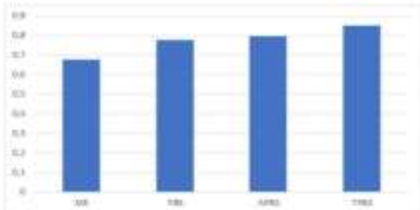


**Fig. 5.** Accuracy Results of Each Method

The performance was also evaluated by calculating the precision, recall, and F-score values for each extraction method used, and the proposed TFBS method was observed to have the highest values while ME had the lowest.

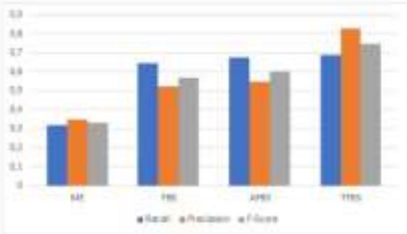


**Fig. 6.** Result of the Recall, Precision, and F-Score measurements for each method

The next experiment was to classify public reactions during the COVID-19 pandemic using the TFBS method with a focus on happy, sad, fear, angry, and love categories. It was discovered from the crawling analysis in Figure 7 that public reactions associated with "love" were higher than others. On July 24-31, 2022 the reactions to the pandemic were observed to be normal with the assumption that it could not attack Indonesia because of the ongoing vaccination program and the creation of herd immunity. The people thought those that were not disciplined with the implementation of the health protocols and refused vaccination were most susceptible to the infection. The "angry" class was found in second place on July 25-27 2022 due to the increase in COVID-19 cases in some areas that have returned to normal activities including schools that have been closed to implement online learning. It was also discovered that other viruses such as monkeypox were becoming an issue in these communities.

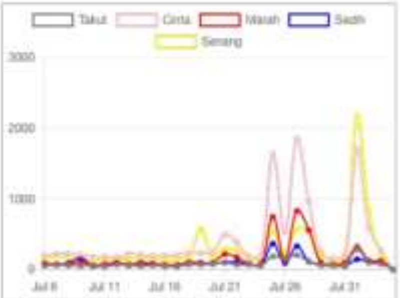



**Fig. 7**. Trend of public reactions from 06/07/2022 to 04/08/2022

Moreover, the trend of public reactions was also visualized in real-time on the dashboard of the monitoring system with a focus on every 2 minutes in the last 1 hour. It was discovered that the trend from 05.20 AM to approximately 06.00 tends more toward fear reactions. This simply indicates the general feeling of the people in the community regarding COVID-19.


**Fig. 8.** Real-time Trend of public reactions in the last 1 hour

The classifications are indicated by visualization in the form of a word cloud as shown in Figures 9 to 13. It was discovered that the words often observed in each class vary. For example, the frequent words in "happy" class include health, protocol, and discipline, those in "sad" are infected, sick, and fever, "love" had health, protocol, and discipline, "fear" recorded results, examination, and ministry of health while "angry" had infected, vaccinated, and positive.


**Fig. 9.** Word Cloud of Popular topic associated with public reactions in the happy category

Figure 9 shows the trend of public reactions in happy category through the word cloud consisting of several frequent words that are widely discussed in the community and formed the opinion of the people. Some of these include protocol, health, masks, and discipline that were trending because, at the time, the enforcement of health protocol discipline was the main thing and had become an issue of public awareness.


**Fig. 10.** Word Cloud of Popular topic associated with public reactions in the sad category

Figure 10 indicates the public reactions for the sad category are dominated by frequent words such as infected, sick, fever, positive, and cough. They were appearing often at the time because work and school activities of people were conducted offline. This led to several positive cases and other symptoms of COVID-19.

Figure 11 shows that the frequent words in the "love" category include health, protocol, discipline, progress, increase, deltacron, variant, and mask. These are almost the same as those observed in the happy category but the love aspect focuses on the hope for the better. During this period, new variants of COVID-19 began to appear such as omicron and deltacron but the public expresses their love while maintaining health protocols and discipline through the usage of masks.


**Fig. 11.** Word Cloud of Popular topic associated with public reactions in the love category

Figure 12 shows that the words trending in the fear category are dominated by the ministry of health, health, laboratory, examination, discipline, and protocol. The focus was placed more on the issue of an increase in COVID-19 patients, the emergence of new variants and the monkeypox virus during the period, and the need for the Ministry of Health to respond through appropriate policies. The public was always vigilant, afraid, and kept reminding each other to always be disciplined in maintaining health protocols.

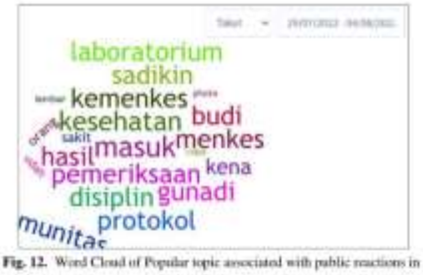



**Fig. 12.** Word Cloud of Popular topic associated with public reactions in the fear category

The trending words related to the fear category were found in Figure 13 to be dominated by words such as sick, affected, positive, death, and vaccine. The cause of these reactions was observed to be almost the same as the sad category.


**Fig. 13.** Word Cloud of Popular topic associated with public reactions in the angry category

**Table 3.** Word Cloud of Popular Topics from Indonesian User

| Label | Original World Cloud | English of World Cloud |
|---|---|---|
| Happy | "vaksinasi, pemalsuan, adhanom, masker angka, kesehtan, test, sosial, ulang, tedros, direktur, protokol, masuk, desak, disiplin, hasil, perkuatjagaimunitas, record menerapkan" | "vaccination, forgery, adhanom, figure mask, health, test, social, repeat, tedros, director, protocol, admission, urge, discipline, results, strengthen immunity, apply record " |
| Sad | "sekolah, batuk, sakit, orang, udah, bener, makan, demam, sedih, anak konser, positif" | "school, cough, sick, people, already, right, eat, get sick, fever, sad, concert children, positive " |
| love | "Omicron, kenaikan, disiplin, protokol, masker, perkuatjagaimunitas, kesehatan, varian, naik, deltacron, prokes, subvarian" | "Omicron, increase, discipline, protocol, mask, strengthen immunity, health, variant, up, deltacron, health protocols, sub-variant " |
| Fear | "Laboratorium, kemenkes, kesehatan, pemeriksaan, sakit, masuk, hasil, kena, disiplin, protokol" | "Laboratory, Ministry of Health, health, examination, illness, admission, results, infected, discipline, protocol " |
| Angry | "Sakit, banget, viral, kena, positif, kematian, vaksin, jumlah" | "Sick, really, viral, infected, positive, death, vaccine, number " |

Table 3 shows some words in the word cloud from each category that are believed to have appeared due to their high frequency of occurrence in text tweets written by users. This means they reflect the opinion developed in the community.

Figure 14 shows a visualization of the report from the COVID-19 public reaction monitoring system with a focus on the tweet recapitulation of each reaction category. It also indicated the graph of people's reactions in real-time which can also be determined at a certain time.


**Fig. 14.** Visualization of the COVID-19 public reaction monitoring system report

## 5. Conclusion

This research developed a sentiment analysis model using the TF-IDF and lexical methods to identify public reactions to COVID-19. The prediction accuracy measurement showed that the proposed method was more accurate than the others due to its value of 0.854. It was also discovered from the Recall, Precision, and F-Score evaluation that the ME method had poor performance while the proposed TFBS had the best followed by AFBS and FBS. The TFBS also ranked higher compared to the other methods because it correctly identified and classified tweets. Moreover, the experiments conducted to classify the tendency of public reactions in dealing with the COVID-19 pandemic using the method showed the possibility of identifying and visualizing several categories on the dashboard of the community reaction monitoring system.


## Acknowledgments

This research was supported by both The Directorate of Research and Community Service – Ministry of Research, Technology and Higher Education and the Semarang State University (UNNES).


## References


[1] R. P. Pratama and A. Tjahyanto, "The influence of fake accounts on sentiment analysis related to COVID-19 in Indonesia," in *Procedia Computer Science*, 2021, vol. 197, no. 2021, pp. 143–150. doi: 10.1016/j.procs.2021.12.128.

[2] A. D. Fakhira, A. S. Pawitra, K. C. Diyanah, E. A. Wikurendra, I. Nagy, and N. S. M. Abdeljawad, "Awareness of Doing 3M (Wearing Mask, Physical Distancing, Washing Hands) During Pandemic Era in Rural and Urban Families," *Jurnal Kesehatan Lingkungan*, vol. 13, no. 2, p. 94, 2021, doi: 10.20473/jkl.v13i2.2021.94-101.

[3] A. K. and M.-V. Dorothée Allain-Dupré, Isabelle Chatry and Michalun, "The territorial impact of COVID-19 : Managing The Crisis Across Levels of Government," 2020.

[4] L. Nahar, Z. Sultana, N. Iqbal, and A. Chowdhury, "Sentiment Analysis and Emotion Extraction: A Review of Research Paradigm," *1st International Conference on Advances in Science, Engineering and Robotics Technology 2019, ICASERT 2019*, no. December, 2019, doi: 10.1109/ICASERT.2019.8934654.

[5] Z. Drus and H. Khalid, "Sentiment analysis in social media and its application: Systematic literature review," *Procedia Computer Science*, vol. 161, pp. 707–714, 2019, doi: 10.1016/j.procs.2019.11.174.

[6] A. Pertiwi, A. Triayudi, and E. T. E. Handayani, "Sentiment Analysis of the Impact of Covid-19 on Indonesia's Economy through Social Media Using the ANN Method," *Jurnal Mantik*, vol. 4, no. 1, pp. 605–612, 2020.

[7] P. Yadav and D. Pandya, "SentiReview: Sentiment Analysis based on Text and Emoticons," in *International Conference on Innovative Mechanisms for Industry Applications*, 2017, pp. 467–472.

S. Peng, G. Wang, and D. Xie, "Social Influence Analysis in Social Networking Big Data: Opportunities and Challenges," *IEEE Network*, vol. 31, no. 1, pp. 11–17, 2017, doi: 10.1109/MNET.2016.1500104NM.

A. Go, L. Huang, and R. Bhayani, "Twitter Sentiment Analysis," 2009.

C. R. Machuca, C. Gallardo, and R. M. Toasa, "Twitter sentiment analysis on coronavirus: Machine learning approach," *Journal of Physics: Conference Series*, vol. 1828, no. 1, 2021, doi: 10.1088/1742-6596/1828/1/012104.

N. N. Alabid and Z. D. Katheeth, "Sentiment analysis of twitter posts related to the covid-19 vaccines," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 24, no. 3, pp. 1727–1734, 2021, doi: 10.11591/ijeecs.v24.i3.pp1727-1734.

I. Srikanth, A. Damodaram, Y. Teekaraman, R. Kuppusamy, and A. R. Thelkar, "Sentiment Analysis on COVID-19 Twitter Data Streams Using Deep Belief Neural Networks," *Computational Intelligence and Neuroscience*, 2022, doi: https://doi.org/10.1155/2022/8898100.

D. A. Efiliandu, E. N. Dianti, and O. G. Khoirunnisa, "Analysis of twitter sentiment in COVID-19 era using fuzzy logic method," *Journal of Soft Computing Exploration*, vol. 1, no. 1, pp. 1-5, 2021, doi: 10.52465/joscex.v2i1.12.

[14] T. Hendrawati and C. P. Yanti, "Analysis of Twitter Users Sentiment against the Covid-19 Outbreak Using the Backpropagation Method with Adam Optimization," *Journal of Electrical, Electronics and Informatics*, vol. 5, no. 1, p. 1, 2021, doi: 10.24843/jeei.2021.v05.i01.p01.

[15] T. N. Simanjuntak and S. Pramana, "Sentiment Analysis on Overseas Tweets on the Impact of COVID-19 in Indonesia," *Indonesian Journal of Statistics and Its Applications*, vol. 5, no. 2, pp. 304–313, 2021, doi: 10.29244/ijsa.v5i2p304-313.

[16] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, 2022, doi: https://doi.org/10.1007/s10462-022-10144-1.

[17] T. N. Prakash and A. Aloysius, "Data Preprocessing in Sentiment Analysis using Twitter Data," *International Education Applied Research Journal*, vol. 3, no. 7, 2019.

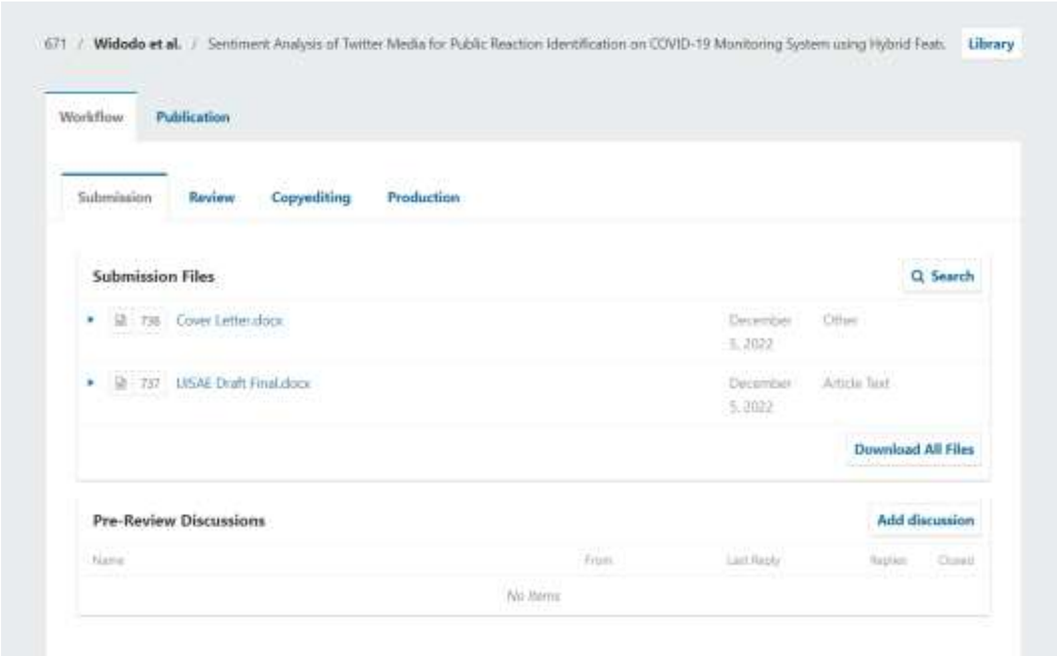
[18] A. F. Hidayatullah and M. R. Ma'arif, "Pre-processing Tasks in Indonesian Twitter Messages," *Journal of Physics: Conference Series, Volume 801, International Conference on Computing and Applied Informatics - IOP Publishing Ltd*, 2016, doi: 10.1088/1742-6596/801/1/012072.

[19] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," *Procedia Computer Science*, vol. 152, pp. 341–348, 2019, doi: 10.1016/j.procs.2019.05.008.

[20] Kasmi and M. Mashuri, "Sentiment Analysis In Twitter Using Lexicon Based and Polarity Multiplication," 2019. doi: 10.1109/ICAIIT.2019.8834477.

[21] N. Medagoda, S. Shanmuganathan, and J. L. Whalley, "Sentiment Lexicon Construction," 2015.

7. **Submit ulang revisi manuskrip dan *cover letter* sebagai berkas pendukung.**



**Editor in Chief**

**International Journal of Intelligent Systems and Application in Engineering**

**Subject : Submit an article to IJISAE Journal**

Dear Editor,

I am enclosing here with a manuscript entitled "Sentiment Analysis of Twitter Media for Public Reaction Identification on COVID-19 Monitoring System using Hybrid Feature Extraction Method" for possible publication in IJISAE.

The purpose of this research article is to develop a sentiment analysis model using TF-IDF and lexical method to identify public reactions due to COVID19.

This manuscript has not been previously published, is not currently submitted for evaluation to any other journal, and will not be submitted elsewhere before a decision is made by this journal.

I would be very grateful if the manuscript to be published in your journal. Look forward to your favorable consideration.

Yours truly,

On behalf Authors,

Djoko Adi Widodo

International Journal of
**INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING**

ISSN:2147-6799 www.ijisae.org Original Research Paper

# Sentiment Analysis of Twitter Media for Public Reaction Identification on COVID-19 Monitoring System using Hybrid Feature Extraction Method

**Djoko Adi Widodo*[1], Nur Iksan[1], Budi Sunarko[1]**

***Abstract:*** Several strategies were implemented to prevent COVID-19 spread. However, these steps were not effectively implemented in the community due to low public awareness and lack of discipline in daily life and this indicated a potential threat of continuous exposure to the virus. It was also observed that the pandemic greatly affected other areas besides the health sector ranging from the social, political, religious, and economic aspects to the resilience of the people. These can be observed through direct observation of the community or activities of the people on social media, especially in relation to the socio-economic aspect. Therefore, this research was conducted using social media, specifically Twitter, via the Twitter API to obtain data related to COVID-19 pandemic in Indonesia. In this research, a sentiment analysis method was developed in this research to identify public opinion related to the spread of the COVID-19 virus and its social impact on society. This was achieved using the TF-IDF and Lexical methods for feature extraction and Naive Bayes for classification. This research used a dataset obtained through Twitter using the keyword "COVID-19" in Indonesian and manually labelled using 5 categories of reactions i.e., fear, angry, love, sad, and happy. The prediction accuracy values showed that the proposed TFBS method had a higher accuracy value of 0.85 compared to other methods. The performance was also evaluated by calculating the precision, recall, and F-score values for each extraction method used, and the proposed TFBS method was observed to have the highest values while ME had the lowest.

***Keywords:*** *Sentiment Analysis, COVID-19, TF-IDF, Lexicon Based*

## 1. Introduction

World Health Organization (WHO) declared the COVID-19 virus a pandemic due to its spread to all countries in the world, thereby, becoming a serious concern for the public, government, and world health institutions [1]. The virus started emerging in Indonesia in early January 2020 and the data from the handling team showed an increasing trend with nearly 600,000 people reported to be exposed and approximately 18,000 died. Several strategies were implemented to prevent further spread such as washing hands, wearing masks, quarantine, social distancing, working from home, and others [2]. However, these steps were not effectively implemented in the community due to low public awareness and lack of discipline in daily life and this indicated a potential threat of continuous exposure to the virus. This led the government to introduce another strategy which involved launching a vaccination program for all elements of society to prevent infection.

This vaccination program was one of the effective efforts made to suppress the spread with the main goal of providing immunity against COVID-19 virus infection and ensuring the formation of herd or group immunity. Furthermore, the spread was also prevented by providing information to increase public awareness on the best way to conduct their activities during the period.

[1] *Department of Electrical Engineering, Faculty of Engineering, Semarang State University – Indonesia*
[2] *Department of Electrical Engineering, Faculty of Engineering, Semarang State University – Indonesia*
[3] *Department of Electrical Engineering, Faculty of Engineering, Semarang State University – Indonesia*
* *Corresponding Author Email: djokoadiwidodo@mail.unnes.ac.id*

In addition to the health sector, the pandemic also significantly affected the economic, religious, security, resilience, political, and social sectors [3]. These were not identified directly as the case with the health sector but through the observation of community activities using different media such as Instagram, Facebook, Twitter, and others. These media serve as social sensors for communities through the uploading of ideas, thoughts, and opinions in the forms of text, video, photos, and audio. The contents are focused on activities and conditions of the things the people feel and see in their surrounding environment which are subsequently sha red on social media platforms. Meanwhile, sentiment analysis can be used to identify and predict future conditions using the information [4–6]. A sentiment analysis method was developed in this research to identify public opinion related to the spread of the COVID-19 virus and its social impact on society [7–9]. This was achieved using the TF-IDF and Lexical methods for feature extraction and Naive Bayes for classification.

## 2. Related Works

Sentiment analysis has become an interesting topic and its methods have been widely developed, especially on Twitter concerning the COVID-19 spread. It was used by Machuca [10] to determine negative and positive sentiments through Machine Learning (ML) and Natural Language Processing (NLP) methods. The Logistic Regression Algorithm was used as an ML algorithm to classify feelings or emotions regarding the COVID-19 topic on Twitter in 2020. Meanwhile, the TF-IDF method was used to extract NLP features and the classification accuracy value was found to be

78.5%. Alabid & Katheeth [11] also developed a model framework to classify public sentiment and opinion regarding the implementation of the COVID-19 vaccination program. The research further compared 2 ML methods including Support Vector Machine (SVM) and Naive Bayes (NB) to classify sentiment polarity into positive, negative, and neutral. The experimental results showed that the NB method had a higher accuracy performance than SVM with 0.81 and 0.75 respectively. Srikanth et al., [12] also conducted a sentiment analysis using a combination of preprocessing methods and word embedding to perform feature extraction. The Deep Belief Neural Network (DBN) algorithm was used to classify tweets through a pseudo-labeling process and the experimental results showed that the DBN algorithm achieved an accuracy value of 90.3% which is better than the other methods. Moreover, Efrilianda [13] developed a sentiment analysis method to determine positive, negative, and neutral polarity using the Backpropagation Multi- Layer Perceptron (MLP) algorithm and the results showed an accuracy value of 70%. It was observed that these studies did not consider the development of a feature extraction method as an important factor in increasing the accuracy of sentiment classification. They also only classified sentiments into negative, positive, and neutral polarities instead of emotions and feelings. Therefore, this research developed a feature extraction method by combining TF-IDF with the lexical method. Sentiment analysis was also developed concerning the community's reactions to the COVID-19 pandemic with a focus on happy, sad, angry, fear, and love categories.

## 3. Method

Sentiment analysis is also known as opinion mining and it is normally used to process natural discussion or text commonly referred to as NLP to determine the sentiments embedded in an opinion. It usually categorizes the opinion sentences in the form of polarity which include Negative, Positive, and Neutral [9, 14]. It can also be used for classification in other forms such as happy, sad, angry, fear, and love. The steps in the sentiment analysis process are shown in the following Figure 1:

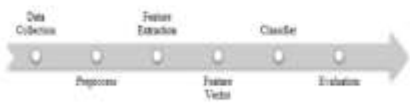


**Fig. 1**. The steps in the sentiment analysis process

The sentiment analysis consists of 6 stages of data collection, data pre-processing, feature extraction, vectorization, classification, and evaluation. The data collection stage was used to retrieve text data from Twitter through the API [15]. The data pre-processing was used to prepare the collected data for extraction into features [16] and this is important to determine the performance of the sentiment analysis process later. Overall, this stage is divided into two which include general pre-processing associated with the text contained in the tweet and special pre-processing related to the characteristics of the tweet text [17]. Figure 2 shows the stages of data pre-processing conducted in this research.


**Fig. 2.** Pre-processing stages

1. Removal URL by filtering the tweet text to remove URLs such as "HTTP://".
2. Removal of the special symbol in the tweets such as hashtags (#), retweets (RT), and usernames (@).
3. Tokenization such that the tweet text in the form of a sentence was be divided into several constituent words or phrases called tokens using the tokenize() function. The function was obtained from the RegexpTokenizer() package in the NLTK library.
4. Case folding which involved changing some words from the tokenization process to lowercase letters using the lower() function.
5. Nonstandard word handling which involved returning the detected unstandardized words such as errors in the use of spelling, abbreviations, slang words, and elongated words to standard form.
6. The stemming process plays an important role in finding the root word of a derivative word. This process will remove affixes to a word. The method used in this process will be different in each language. In Indonesian usage, the stemming process uses the Literature library, while in English it uses the NLTK library.
7. Stopword Removal stage which involved removing common words that do not have a significant effect on the tweet sentence. This was achieved through the Literature and NLTK libraries.

The next stage is extraction which is very fundamental in the sentiment analysis process [18]. It involves converting the pre-processed text into vector form and was conducted in this research using the TF-IDF approach which normally extracts sentence features based on the frequency of word occurrence. Moreover, the frequency term was calculated by determining a feature based on the consistent appearance of a word. The features represent a document and have a high weight. In contrast to TF-IDF, the frequency of a word occurrence in a document is usually compared with an entire database. In a situation the word appears too often in several documents, it cannot be used as a feature because it is considered a general word that does not represent a specific document. The TF-IDF weight was calculated using the following Equations (1) and (2):

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

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

The lexical method was also applied in this research and it involves using the word meaning to determine the sentence features. The accuracy of the sentiment analysis is higher when the word meaning is more complete [19, 20]. The use of this method usually focuses on applying word lists such as "clean", "healthy", "good", and "safe" to determine positive polarity and others such as "sloppy", "bad", "and slow" to determine the negative polarity. This feature extraction for the English lexicon was conducted using the library from the Google Translate API and the lexicon was translated first after which the sentiment scores were calculated from sentiwordnet using Equations (3) and (4).

$$Score = \sum \frac{(PosScore - NegScore)}{tot_{index}} \quad (3)$$

$$Sentence_{Score} = \sum Score \quad (4)$$

Meanwhile, the feature extraction of the Indonesian language lexicon was conducted by calculating the appearance of the word opinion in each sentence after which the positive and negative sentiments were determined using Equations (5) and (6).

$$Score = \sum frek_{opini_{word}} \quad (5)$$

$$Sentence_{Score} = \sum_{i=0}^{n} \sum Score_i \quad (6)$$

This means TF-IDF and Lexical methods which are statistical and semantic approaches respectively were combined in this research to determine the weighting of the feature extraction process. The proposed sentiment analysis model is presented in the following Figure 3.

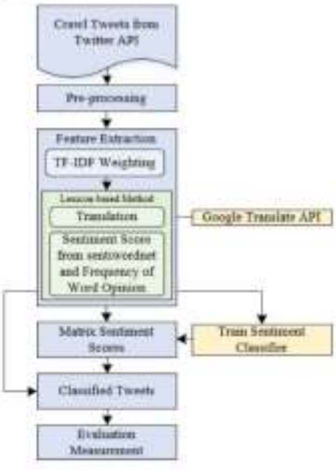



**Fig. 3.** Proposed Sentiment Analysis Model

At the classification stage, the Naive Bayes algorithm was used to determine the tweets consisting of 5 classes which include fear, angry, love, sad, and happy. Furthermore, the performance evaluation was conducted by comparing the Feature-Based Sentiment algorithm using TF-IDF and Lexicon (TFBS) with several other methods such as Maximum Entropy (ME), Feature-Based Sentiment (FBS), and Augmented Feature-Based Sentiment (AFBS).

Performance evaluation was conducted by calculating the accuracy, precision, recall, and F-Score. The accuracy value was calculated by comparing the number of correct data and the total amount of data. A higher accuracy value indicates a better method. Moreover, precision, recall, and F-score were measured to determine any deviations in the data. Precision value was determined by comparing the number of relevant classification data with the total amount of data from a particular class. The recall value was obtained from the comparison of the amount of relevant data with the total relevant data while the F-Score value was determined by calculating the average value of precision and recall. The equations used to calculate precision, recall, and f-score values are stated as Equations (7), (8) and (9).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (7)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (8)$$

$$F - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (9)$$

## 4. Result and Discussion

This research used a dataset obtained through Twitter using the keyword "COVID-19" in Indonesian. The dataset was retrieved through a crawling process using the Tweepy library in Python for as many as 3148 lines. It was subsequently passed through the normalization stage to eliminate duplicated data such as the tweets reposted by other users or retweets (RT). The dataset was manually labelled using 5 categories of reactions as shown in Table 1.

**Table 1.** Labelled Dataset using 5 categories of reactions

| Text Tweet (Original) | Text Tweet (after translated in English) | Labels |
|---|---|---|
| Positif covid-19 bukan kutukan dan bukan aib. Terima kasih sudah berbagi dan tetap semangat. | Positive covid-19 is not a curse and not a disgrace. Thank you for sharing and keep the spirit. | Happy |
| sedih banget kalo denger cerita perawat yang jadi satgas covid-19. banyak pasien covid yang bandel dan susah dibilangin ternyata | It's really sad to hear the story of the nurse who became the Covid-19 task force. many covid patients are stubborn and hard to say it turns out | Sad |
| gak respect banget.. ini ada sodara meninggal di RS padahal karena riwayat stroke, udah dimandiin dan diangkat pas beres malah keluar surat positif COVID-19?! Sumpah baru pertama kali ngeliat langsung kasus orang meninggal dicovidkan dan | really no respect.. this is a friend who died in the hospital even though because of a history of stroke, he was washed and picked up when it was done, instead a positive letter for COVID-19 came out?! I swear it's the first | Angry |

| Text Tweet (Original) | Text Tweet (after translated in English) | Labels |
|---|---|---|
| bener adanya 😢😢 | time I've seen a case of someone who died being COVID-19 and it's true | |
| Besok mau di vaksin covid :( Harusnya seneng ga sih Kan nntinya dpt perlindungan, jujur malah takut :( | Tomorrow I'm going to be vaccinated against covid :( Should I be happy or not, I'm going to get protection, to be honest I'm scared :( | Fear |
| Rindu suasana belajar mengajar sebelum covid mewabah di negri ini | Missing the teaching and learning atmosphere before the covid outbreak in this country | Love |


**Fig. 4.** Example of a text tweet on Twitter

The tweets contain some useless information or noise such as hashtags (#), Html (http://www), mentions (@), numbers, local language usage, and abbreviations. The initial stage in the sentiment analysis process is the pre-processing which involves removing the URL, special symbol, punctuation, tokenization, case folding, and stemming as shown in Figure 2. The useless information was removed by omitting URLs, special symbols, ASCII numbers, and punctuations. Figure 4 shows an example of a tweet containing noise and this means there is a need for pre-processing. The tweet text in Figure 4 was pre-processed and the results are presented in the following Table 2.

**Table 2.** Cleansing at the Pre-processing Stage

| Preprocessing stage | Original Tweet | Tweet in English |
|---|---|---|
| Tweet | " Orang-orang begitu mana paham covid-19, dijelasin berulangpun tetep ngeyel. What should i do? . . . Yaudah biarin aja,kalo positif apalagi sampe bergejala ga akan diurus. Simpan tenagamu untuk yang benar-benar membutuhkan ya nakes 😩. Kita kuat kita bisa survive" | " How do people understand covid-19, even when they explain it again and again, it's still annoying. What should I do? . . . Well, just let it be, if it's positive, especially if you have symptoms, it won't be taken care of. Save your energy for those who really need it, guys. We are strong we can survive" |
| Preprocessing Results | "orang paham covid jelas ulang udah biar saja positif gejala urus simpan tenaga butuh ya nakes kuat survive" | "people who understand that covid is being explained again, they keep yelling, what should i do. let's just leave it if it's positive until the symptoms don't take care of saving energy, we need strong health workers to survive" |
| Tokenization Result | [orang] [paham] [covid] [jelas] [ulang] [udah] [biar] [saja] [positif] [gejala] [urus] [simpan] [tenaga] [butuh] [nakes] [kuat] [survive] | [people] [understand] [covid] [clear] [repeat] [already] [let it be] [just] [positive] [symptoms] [manage] [save] [energy] [need] [nakes] [strong] [survival] |

Several processes were implemented at the pre-processing stage and these include the deletion of emoticons, URLs, special symbols, ASCII symbols, and punctuation marks. Furthermore, the text characteristics were handled using different methods such as tokenization, case folding, non-standard word handling, stemming, and stop-word removal. The results were later forwarded to the feature extraction stage to produce a scoring matrix by combining the TF-IDF and lexicon methods.

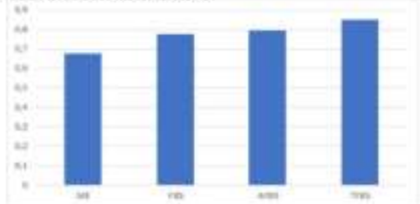


**Fig. 5.** The accuracy values obtained from using different feature extraction methods

Figure 5 shows the accuracy values obtained from using different feature extraction methods. It is important to note that the focus was placed on the tweet classification using five classes of fear, angry, love, sad, and happy. The results showed that the proposed TFBS method had a higher accuracy value of 0.85 compared to other methods.

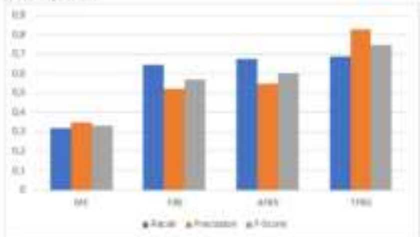


**Fig. 6.** Result of the Recall, Precision, and F-Score measurements for each method

As shown in Figure 6, the performance was also evaluated by calculating the precision, recall, and F-score values for each extraction method used, and the proposed TFBS method was observed to have the highest values while ME had the lowest.

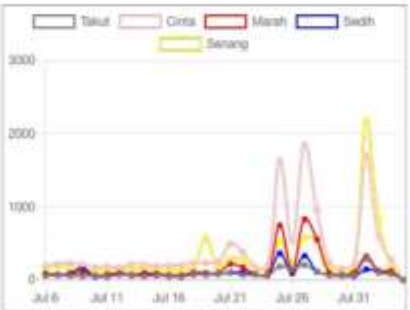



**Fig. 7.** Trend of public reactions in 06 July 2022 to 04 Aug 2022

The next experiment was to classify public reactions during the COVID-19 pandemic using the TFBS method with a focus on happy, sad, fear, angry, and love categories. It was discovered from the crawling analysis in Figure 7 that public reactions associated with "love" were higher than others. On July 24-31, 2022 the reactions to the pandemic were observed to be normal with the assumption that it could not attack Indonesia because of the ongoing vaccination program and the creation of herd immunity. The people thought those that were not disciplined with the implementation of the health protocols and refused vaccination were most susceptible to the infection. The "angry" class was found in second place on July 25-27 2022 due to the increase in COVID-19 cases in some areas that have returned to normal activities including schools that have been closed to implement online learning. It was also discovered that other viruses such as monkeypox were becoming an issue in these communities.

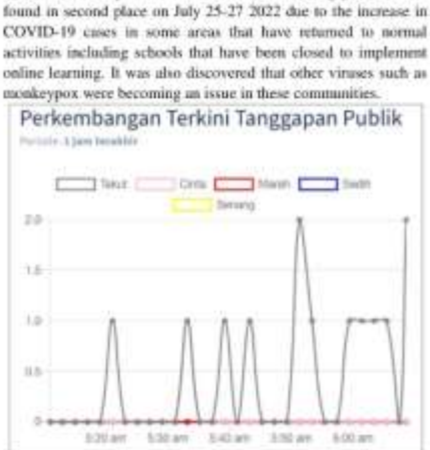



**Fig. 8.** Real- time on the dashboard of the monitoring system every 2 minutes in the last 1 hour

Moreover, in Figure 8, the trend of public reactions was also visualized in real- time on the dashboard of the monitoring system with a focus on every 2 minutes in the last 1 hour. It was discovered that the trend from 05.20 AM to approximately 06.00 tends more toward fear reactions. This simply indicates the general feeling of the people in the community regarding COVID-19.

The classifications are indicated by visualization in the form of a word cloud as shown in Figures 9 to 13. It was discovered that the words often observed in each class vary. For example, the frequent words in "happy" class include health, protocol, and discipline, those in "sad" are infected, sick, and fever, "love" had health, protocol, and discipline, "fear" recorded results, examination, and ministry of health while "angry" had infected, vaccinated, and positive.


**Fig. 9.** Popular topic of Word Cloud that associated with happy reaction

Figure 9 shows the trend of public reactions in happy category through the word cloud consisting of several frequent words that are widely discussed in the community and formed the opinion of the people. Some of these include protocol, health, masks, and discipline that were trending because, at the time, the enforcement of health protocol discipline was the main thing and had become an issue of public awareness.


**Fig. 10.** Popular topic of Word Cloud that associated with sad reaction

Figure 10 indicates the public reactions for the sad category are dominated by frequent words such as infected, sick, fever, positive, and cough. They were appearing often at the time because work and school activities of people were conducted offline. This led to several positive cases and other symptoms of COVID-19.


**Fig. 11.** Popular topic of Word Cloud that associated with love reaction

Figure 11 shows that the frequent words in the "love" category include health, protocol, discipline, progress, increase, deltacron, variant, and mask. These are almost the same as those observed in the happy category but the love aspect focuses on the hope for the better. During this period, new variants of COVID-19 began to appear such as omicron and deltacron but the public expresses their love while maintaining health protocols and discipline through the usage of masks.


Fig. 12. Popular topic of Word Cloud that associated with fear reaction

Figure 12 shows that the words trending in the fear category are dominated by the ministry of health, health, laboratory, examination, discipline, and protocol. The focus was placed more on the issue of an increase in COVID-19 patients, the emergence of new variants and the monkeypox virus during the period, and the need for the Ministry of Health to respond through appropriate policies. The public was always vigilant, afraid, and kept reminding each other to always be disciplined in maintaining health protocols.


Fig. 13. Popular topic of Word Cloud that associated with angry reaction

The trending words related to the fear category were found in Figure 13 to be dominated by words such as sick, affected, positive, death, and vaccine. The cause of these reactions was observed to be almost the same as the sad category.

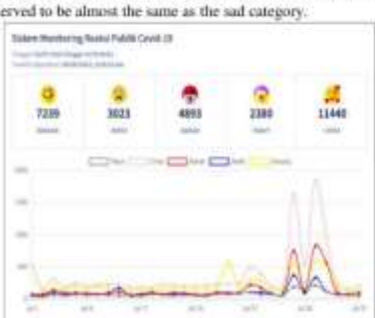



Fig. 14. Visualization of the COVID-19 public reaction monitoring system report

Some words in the word cloud from each category that are believed to have appeared due to their high frequency of occurrence in text tweets written by users. This means they reflect the opinion developed in the community. Figure 14 shows a visualization of the report from the COVID-19 public reaction monitoring system with a focus on the tweet recapitulation of each reaction category. It also indicated the graph of people's reactions in real-time which can also be determined at a certain time.

## 5. Conclusion

This research developed a sentiment analysis model using the TF-IDF and lexical methods to identify public reactions to COVID-19. The prediction accuracy measurement showed that the proposed method was more accurate than the others due to its value of 0.854. It was also discovered from the Recall, Precision, and F-Score evaluation that the ME method had poor performance while the proposed TFBS had the best followed by AFBS and FBS. The TFBS also ranked higher compared to the other methods because it correctly identified and classified tweets. Moreover, the experiments conducted to classify the tendency of public reactions in dealing with the COVID-19 pandemic using the method showed the possibility of identifying and visualizing several categories on the dashboard of the community reaction monitoring system.

## References

[1] R. P. Pratama and A. Tjahyanto, "The influence of fake accounts on sentiment analysis related to COVID-19 in Indonesia," in Procedia Computer Science, 2021, vol. 197, no. 2021, pp. 143–150, doi: 10.1016/j.procs.2021.12.128.

[2] A. D. Fakhira, A. S. Pawitra, K. C. Diyanah, E. A. Wikurendra, I. Nagy, and N. S. M. Abdeljawad, "Awareness of Doing 3M (Wearing Mask, Physical Distancing, Washing Hands) During Pandemic Era in Rural and Urban Families," J. Kesehat. Lingkung., vol. 13, no. 2, p. 94, 2021, doi: 10.20473/jkl.v13i2.2021.94-101.

[3] A. K. and M.-V. Dorothée Allain-Dupré, Isabelle Chatry and Michalun, "The territorial impact of COVID-19 : Managing The Crisis Across Levels of Government," 2020. [Online]. Available: https://www.oecd.org/coronavirus/policy-responses/the-territorial-impact-of-covid-19-managing-the-crisis-across-levels-of-government-d3e314e1/.

[4] L. Nahar, Z. Sultana, N. Iqbal, and A. Chowdhury, "Sentiment Analysis and Emotion Extraction: A Review of Research Paradigm," 1st Int. Conf. Adv. Sci. Eng. Robot. Technol. 2019, ICASERT 2019, no. December, 2019, doi: 10.1109/ICASERT.2019.8934654.

[5] Z. Drus and H. Khalid, "Sentiment analysis in social media and its application: Systematic literature review," Procedia Comput. Sci., vol. 161, pp. 707–714, 2019, doi: 10.1016/j.procs.2019.11.174.

[6] A. Pertiwi, A. Triayudi, and E. T. E. Handayani, "Sentiment Analysis of the Impact of Covid-19 on Indonesia's Economy through Social Media Using the ANN Method," J. Mantik, vol. 4, no. 1, pp. 605–612, 2020, [Online]. Available: https://iocscience.org/ejournal/index.php/mantik.

[7] P. Yadav and D. Pandya, "SentiReview: Sentiment Analysis based on Text and Emoticons," in International Conference on Innovative Mechanisms for Industry Applications, 2017, pp. 467–472.

[8] S. Peng, G. Wang, and D. Xie, "Social Influence Analysis in Social Networking Big Data: Opportunities and Challenges,"

IEEE Netw., vol. 31, no. 1, pp. 11–17, 2017, doi: 10.1109/MNET.2016.1500104NM.

[9] A. Go, L. Huang, and R. Bhayani, "Twitter Sentiment Analysis," 2009.

[10] C. R. Machuca, C. Gallardo, and R. M. Toasa, "Twitter sentiment analysis on coronavirus: Machine learning approach," J. Phys. Conf. Ser., vol. 1828, no. 1, 2021, doi: 10.1088/1742-6596/1828/1/012104.

[11] N. N. Alabid and Z. D. Katheeth, "Sentiment analysis of twitter posts related to the covid-19 vaccines," Indones. J. Electr. Eng. Comput. Sci., vol. 24, no. 3, pp. 1727–1734, 2021, doi: 10.11591/ijeecs.v24.i3.pp1727-1734.

[12] J. Srikanth, A. Damodaram, Y. Teekaraman, R. Kuppusamy, and A. R. Thelkar, "Sentiment Analysis on COVID-19 Twitter Data Streams Using Deep Belief Neural Networks," Comput. Intell. Neurosci., 2022, doi: https://doi.org/10.1155/2022/8898100.

[13] D. A. Efrilianda, E. N. Dianti, and O. G. Khoirunnisa, "Analysis of twitter sentiment in COVID-19 era using fuzzy logic method," J. Soft Comput. Explor., vol. 1, no. 1, pp. 1–5, 2021, doi: 10.52465/joscex.v2i1.12.

[14] T. N. Simanjuntak and S. Pramana, "Sentiment Analysis on Overseas Tweets on the Impact of COVID-19 in Indonesia," Indones. J. Stat. Its Appl., vol. 5, no. 2, pp. 304–313, 2021, doi: 10.29244/ijsa.v5i2p304-313.

[15] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," Artif. Intell. Rev., 2022, doi: https://doi.org/10.1007/s10462-022-10144-1.

[16] T. N. Prakash and A. Aloysius, "Data Preprocessing in Sentiment Analysis using Twitter Data," Int. Educ. Appl. Res. J., vol. 3, no. 7, 2019.

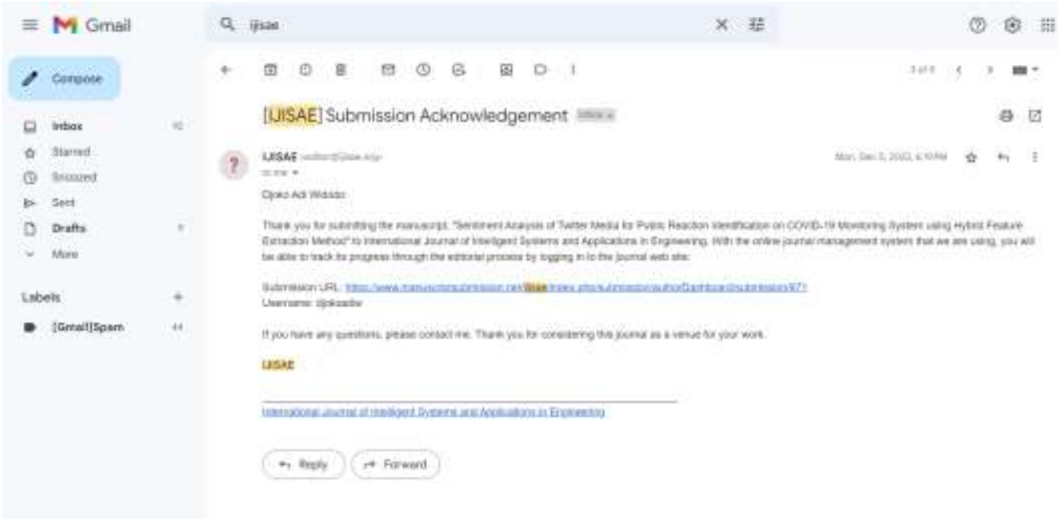
[17] A. F. Hidayatullah and M. R. Ma'arif, "Pre-processing Tasks in Indonesian Twitter Messages," J. Phys. Conf. Ser. Vol. 801, Int. Conf. Comput. Appl. Informatics - IOP Publ. Ltd, 2016, doi: 10.1088/1742-6596/801/1/012072.

[18] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," Procedia Comput. Sci., vol. 152, pp. 341–348, 2019, doi: 10.1016/j.procs.2019.05.008.

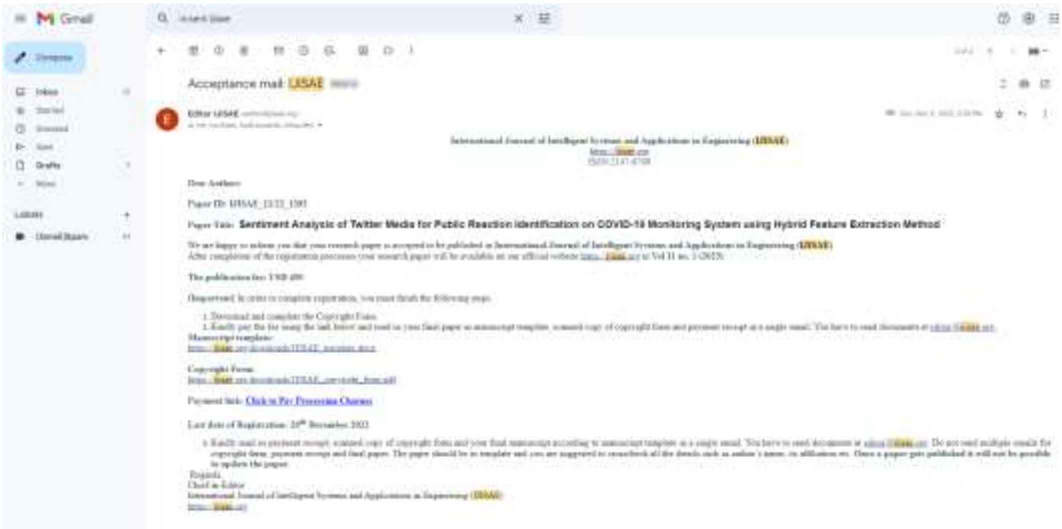
[19] Kusrini and M. Mashuri, "Sentiment Analysis In Twitter Using Lexicon Based and Polarity Multiplication," 2019, doi: 10.1109/ICAIIT.2019.8834477.

[20] [N. Medagoda, S. Shanmuganathan, and J. L. Whalley, "Sentiment Lexicon Construction," 2015, [Online]. Available: 10.1109/ICNC.2015.7378094.

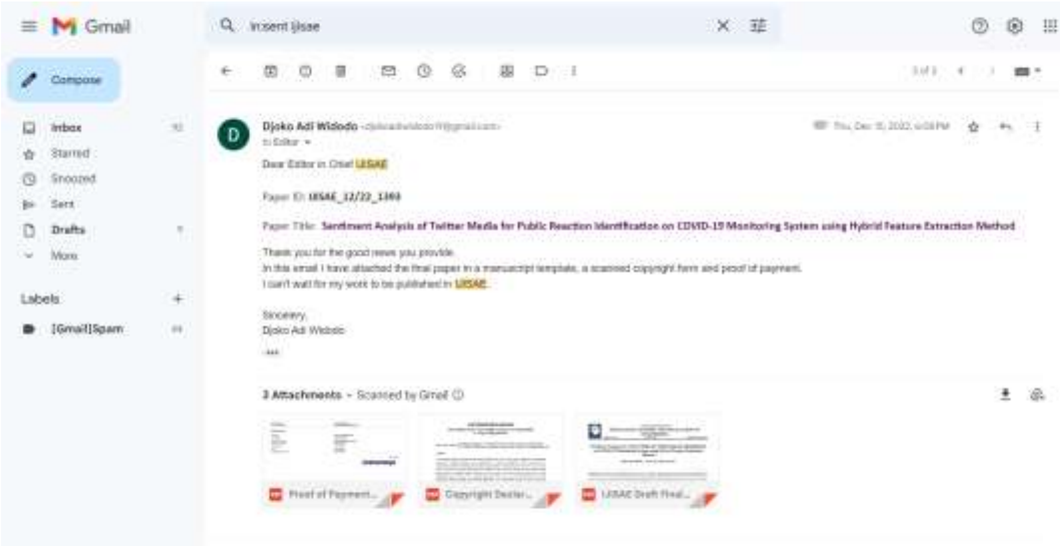
8. **Pemberitahuan via email bahwa manuskrip telah berhasil di submit dan telah diterima.**



9. **Pemberitahuan manuskrip diterima untuk publikasi, permintaan pembayaran *article publication* dan perbaikan manuskrip sesuai template.**



10. **Mengirim manuskrip sesuai template, bukti pembayaran dan *copyright declaration***



International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

IJISAE

ISSN:2147-6799 www.ijisae.org Original Research Paper

# Sentiment Analysis of Twitter Media for Public Reaction Identification on COVID-19 Monitoring System using Hybrid Feature Extraction Method

**Djoko Adi Widodo*[1], Nur Iksan[1], Budi Sunarko[1]**

**Submitted:** 19/10/2022 **Accepted:** 31/12/2022

**Abstract:** Several strategies were implemented to prevent COVID-19 spread. However, these steps were not effectively implemented in the community due to low public awareness and lack of discipline in daily life and this indicated a potential threat of continuous exposure to the virus. It was also observed that the pandemic greatly affected other areas besides the health sector ranging from the social, political, religious, and economic aspects to the resilience of the people. These can be observed through direct observation of the community or activities of the people on social media, especially in relation to the socio-economic aspect. Therefore, this research was conducted using social media, specifically Twitter, via the Twitter API to obtain data related to COVID-19 pandemic in Indonesia. In this research, a sentiment analysis method was developed in this research to identify public opinion related to the spread of the COVID-19 virus and its social impact on society. This was achieved using the TF-IDF and Lexical methods for feature extraction and Naive Bayes for classification. This research used a dataset obtained through Twitter using the keyword "COVID-19" in Indonesian and manually labelled using 5 categories of reactions i.e., fear, angry, love, sad, and happy. The prediction accuracy values showed that the proposed TFBS method had a higher accuracy value of 0.85 compared to other methods. The performance was also evaluated by calculating the precision, recall, and F-score values for each extraction method used, and the proposed TFBS method was observed to have the highest values while ME had the lowest.

**Keywords:** *Sentiment Analysis, COVID-19, TF-IDF, Lexicon Based*

## 1. Introduction

World Health Organization (WHO) declared the COVID-19 virus a pandemic due to its spread to all countries in the world, thereby, becoming a serious concern for the public, government, and world health institutions [1]. The virus started emerging in Indonesia in early January 2020 and the data from the handling team showed an increasing trend with nearly 600,000 people reported to be exposed and approximately 18,000 died. Several strategies were implemented to prevent further spread such as washing hands, wearing masks, quarantine, social distancing, working from home, and others [2]. However, these steps were not effectively implemented in the community due to low public awareness and lack of discipline in daily life and this indicated a potential threat of continuous exposure to the virus. This led the government to introduce another strategy which involved launching a vaccination program for all elements of society to prevent infection.

This vaccination program was one of the effective efforts made to suppress the spread with the main goal of providing immunity against COVID-19 virus infection and ensuring the formation of herd or group immunity. Furthermore, the spread was also prevented by providing information to increase public awareness on the best way to conduct their activities during the period.

In addition to the health sector, the pandemic also significantly affected the economic, religious, security, resilience, political, and social sectors [3]. These were not identified directly as the case with the health sector but through the observation of community activities using different media such as Instagram, Facebook, Twitter, and others. These media serve as social sensors for communities through the uploading of ideas, thoughts, and opinions in the forms of text, video, photos, and audio. The contents are focused on activities and conditions of the things the people feel and see in their surrounding environment which are subsequently sha red on social media platforms. Meanwhile, sentiment analysis can be used to identify and predict future conditions using the information [4–6]. A sentiment analysis method was developed in this research to identify public opinion related to the spread of the COVID-19 virus and its social impact on society [7–9]. This was achieved using the TF-IDF and Lexical methods for feature extraction and Naive Bayes for classification.

[1] *Department of Electrical Engineering, Faculty of Engineering, Semarang State University – Indonesia*
[2] *Department of Electrical Engineering, Faculty of Engineering, Semarang State University – Indonesia*
[3] *Department of Electrical Engineering, Faculty of Engineering, Semarang State University – Indonesia*
* *Corresponding Author Email: djokoadiwidodo@mail.unnes.ac.id*

## 2. Related Works

Sentiment analysis has become an interesting topic and its methods have been widely developed, especially on Twitter concerning the COVID-19 spread. It was used by Machuca [10] to determine negative and positive sentiments through Machine Learning (ML) and Natural Language Processing (NLP) methods. The Logistic Regression Algorithm was used as an ML algorithm to classify feelings or emotions regarding the COVID-19 topic on Twitter in 2020. Meanwhile, the TF-IDF method was used to extract NLP features and the classification accuracy value was found to be 78.5%. Alabid & Katheeth [11] also developed a model framework to classify public sentiment and opinion regarding the implementation of the COVID-19 vaccination program. The research further compared 2 ML methods including Support Vector Machine (SVM) and Naive Bayes (NB) to classify sentiment polarity into positive, negative, and neutral. The experimental results showed that the NB method had a higher accuracy performance than SVM with 0.81 and 0.75 respectively. Srikanth et al., [12] also conducted a sentiment analysis using a combination of preprocessing methods and word embedding to perform feature extraction. The Deep Belief Neural Network (DBN) algorithm was used to classify tweets through a pseudo-labeling process and the experimental results showed that the DBN algorithm achieved an accuracy value of 90.3% which is better than the other methods. Moreover, Efrilianda [13] developed a sentiment analysis method to determine positive, negative, and neutral polarity using the Backpropagation Multi- Layer Perceptron (MLP) algorithm and the results showed an accuracy value of 70%. It was observed that these studies did not consider the development of a feature extraction method as an important factor in increasing the accuracy of sentiment classification. They also only classified sentiments into negative, positive, and neutral polarities instead of emotions and feelings. Therefore, this research developed a feature extraction method by combining TF-IDF with the lexical method. Sentiment analysis was also developed concerning the community's reactions to the COVID-19 pandemic with a focus on happy, sad, angry, fear, and love categories.

## 3. Method

Sentiment analysis is also known as opinion mining and it is normally used to process natural discussion or text commonly referred to as NLP to determine the sentiments embedded in an opinion. It usually categorizes the opinion sentences in the form of polarity which include Negative, Positive, and Neutral [9, 14]. It can also be used for classification in other forms such as happy, sad, angry, fear, and love. The steps in the sentiment analysis process are shown in the following Figure 1:

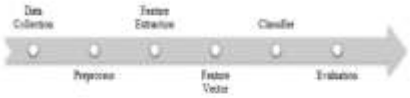


**Fig. 1**. The steps in the sentiment analysis process

The sentiment analysis consists of 6 stages of data collection, data pre-processing, feature extraction, vectorization, classification, and evaluation. The data collection stage was used to retrieve text data from Twitter through the API [15]. The data pre-processing was used to prepare the collected data for extraction into features [16] and this is important to determine the performance of the sentiment analysis process later. Overall, this stage is divided into two which include general pre-processing associated with the text contained in the tweet and special pre-processing related to the characteristics of the tweet text [17]. Figure 2 shows the stages of data pre-processing conducted in this research.


**Fig. 2.** Pre-processing stages

1. Removal URL by filtering the tweet text to remove URLs such as "HTTP://".
2. Removal of the special symbol in the tweets such as hashtags (#), retweets (RT), and usernames (@).
3. Tokenization such that the tweet text in the form of a sentence was be divided into several constituent words or phrases called tokens using the tokenize() function. The function was obtained from the RegexpTokenizer() package in the NLTK library.
4. Case folding which involved changing some words from the tokenization process to lowercase letters

using the lower() function.

5. Nonstandard word handling which involved returning the detected unstandardized words such as errors in the use of spelling, abbreviations, slang words, and elongated words to standard form.
6. The stemming process plays an important role in finding the root word of a derivative word. This process will remove affixes to a word. The method used in this process will be different in each language. In Indonesian usage, the stemming process uses the Literature library, while in English it uses the NLTK library.
7. Stopword Removal stage which involved removing common words that do not have a significant effect on the tweet sentence. This was achieved through the Literature and NLTK libraries.

The next stage is extraction which is very fundamental in the sentiment analysis process [18]. It involves converting the pre- processed text into vector form and was conducted in this research using the TF-IDF approach which normally extracts sentence features based on the frequency of word occurrence. Moreover, the frequency term was calculated by determining a feature based on the consistent appearance of a word. The features represent a document and have a high weight. In contrast to TF-IDF, the frequency of a word occurrence in a document is usually compared with an entire database. In a situation the word appears too often in several documents, it cannot be used as a feature because it is considered a general word that does not represent a specific document. The TF-IDF weight was calculated using the following Equations (1) and (2):

$$tf.idf_{t,d} = tf_{t,d} \times idf_t \tag{1}$$

$$idf_t = \log \frac{N}{df_t} \tag{2}$$

The lexical method was also applied in this research and it involves using the word meaning to determine the sentence features. The accuracy of the sentiment analysis is higher when the word meaning is more complete [19, 20]. The use of this method usually focuses on applying word lists such as "clean", "healthy", "good", and "safe" to determine positive polarity and others such as "sloppy", "bad", "and slow" to determine the negative polarity. This feature extraction for the English lexicon was conducted using the library from the Google Translate API and the lexicon was translated first after which the sentiment scores were calculated from sentiwordnet using Equations (3) and (4).

$$Score = \sum \frac{(PosScore - NegScore)}{tot_{index}} \tag{3}$$

$$Sentence_{Score} = \sum Score \tag{4}$$

Meanwhile, the feature extraction of the Indonesian language lexicon was conducted by calculating the appearance of the word opinion in each sentence after which the positive and negative sentiments were determined using Equations (5) and (6).

$$Score = \sum frek_{opini_{word}} \tag{5}$$

$$Sentence_{Score} = \sum_{i=0}^{n} \sum Score_i \tag{6}$$

This means TF-IDF and Lexical methods which are statistical and semantic approaches respectively were combined in this research to determine the weighting of the feature extraction process. The proposed sentiment analysis model is presented in the following Figure 3.

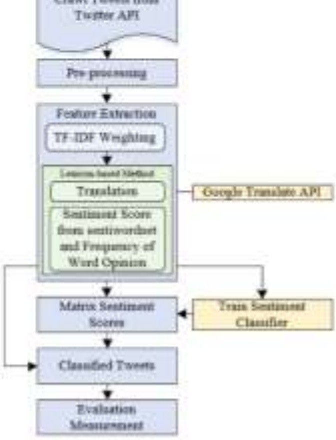



**Fig. 3.** Proposed Sentiment Analysis Model

At the classification stage, the Naive Bayes algorithm was used to determine the tweets consisting of 5 classes which include fear, angry, love, sad, and happy. Furthermore, the performance evaluation was conducted by comparing the Feature-Based Sentiment algorithm using TF-IDF and Lexicon (TFBS) with several other methods such as Maximum Entropy (ME), Feature-Based Sentiment (FBS), and Augmented Feature-Based Sentiment (AFBS).

Performance evaluation was conducted by calculating the accuracy, precision, recall, and F-Score. The accuracy value was calculated by comparing the number of correct data and the total amount of data. A higher accuracy value indicates a better method. Moreover, precision, recall, and F-score were measured to determine any deviations in the data. Precision value was determined by comparing the number of relevant classification data with the total amount of data from a particular class. The recall value was obtained from the comparison of the amount of relevant data with the total relevant data while the F-Score value was determined by calculating the average value of precision and recall. The equations used to calculate precision, recall, and f-score values are stated as Equations (7), (8) and (9).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (7)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (8)$$

$$F - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (9)$$

## 4. Result and Discussion

This research used a dataset obtained through Twitter using the keyword "COVID-19" in Indonesian. The dataset was retrieved through a crawling process using the Tweepy library in Python for as many as 3148 lines. It was subsequently passed through the normalization stage to eliminate duplicated data such as the tweets reposted by other users or retweets (RT). The dataset was manually labelled using 5 categories of reactions as shown in Table 1.

**Table 1.** Labelled Dataset using 5 categories of reactions

| Text Tweet (Original) | Text Tweet (after translated in English) | Labels |
|---|---|---|
| Positif covid-19 bukan kutukan dan bukan aib. Terima kasih sudah berbagi dan tetap semangat. 🎧 | Positive covid-19 is not a curse and not a disgrace. Thank you for sharing and keep the spirit. 🎧 | Happy |
| sedih banget kalo denger cerita perawat yang jadi satgas covid-19. banyak pasien covid yang bandel dan susah dibilangin ternyata | It's really sad to hear the story of the nurse who became the Covid-19 task force. many covid patients are stubborn and hard to say it turns out | Sad |
| gak respect banget.. ini ada sodara meninggal di RS padahal karena riwayat stroke, udah dimandiin dan diangkat pas beres malah keluar surat positif COVID-19?! Sumpah baru pertama kali ngeliat langsung kasus orang meninggal dicovidkan dan bener adanya 😔😔 | really no respect.. this is a friend who died in the hospital even though because of a history of stroke, he was washed and picked up when it was done, instead a positive letter for COVID-19 came out?! I swear it's the first time I've seen a case of someone who died being COVID-19 and it's true | Angry |
| Besok mau di vaksin covid :( Hrusnya seneng ga sih Kan nntinya dpt perlindungan, jujur malah takut :( | Tomorrow I'm going to be vaccinated against covid :( Should I be happy or not, I'm going to get protection, to be honest I'm scared :( | Fear |
| Rindu suasana belajar mengajar sebelum covid mewabah di negri ini | Missing the teaching and learning atmosphere before the covid outbreak in this country | Love |


**Fig. 4.** Example of a text tweet on Twitter

The tweets contain some useless information or noise such as hashtags (#), Html (http://www), mentions (@), numbers, local language usage, and abbreviations. The initial stage in the sentiment analysis process is the pre-processing which involves removing the URL, special symbol, punctuation, tokenization, case folding, and stemming as shown in Figure 2. The useless information was removed by omitting URLs, special symbols, ASCII numbers, and punctuations. Figure 4 shows an example of a tweet containing noise and this means there is a need for pre- processing. The tweet text in Figure 4 was pre-processed and the results are presented in the following Table 2.

**Table 2.** Cleansing at the Pre-processing Stage

| Preprocessing stage | Original Tweet | Tweet in English |
|---|---|---|
| Tweet | " Orang-orang begitu mana paham covid-19, dijelasin berulangpun tetep ngeyel. What should i do? . . . Yaudah biarin aja,kalo positif apalagi sampe bergejala ga akan diurus. Simpan tenagamu untuk yang benar-benar membutuhkan ya nakes 😢. Kita kuat kita bisa survive" | " How do people understand covid-19, even when they explain it again and again, it's still annoying. What should I do? . . . Well, just let it be, if it's positive, especially if you have symptoms, it won't be taken care of. Save your energy for those who really need it, guys. We are strong we can survive" |
| Preprocessing Results | "orang paham covid jelas ulang udah biar saja positif gejala urus simpan tenaga butuh ya nakes kuat survive" | "people who understand that covid is being explained again, they keep yelling, what should i do, let's just leave it if it's positive until the symptoms don't take care of saving energy, we need strong health workers to survive" |
| Tokenization Result | [orang] [paham] [covid] [jelas] [ulang] [udah] [biar] [saja] [positif] [gejala] [urus] [simpan] [tenaga] [butuh] [nakes] [kuat] [survive] | [people] [understand] [covid] [clear] [repeat] [already] [let it be] [just] [positive] [symptoms] [manage] [save] [energy] [need] [nakes] [strong] [survival] |

Several processes were implemented at the pre-processing stage and these include the deletion of emoticons, URLs, special symbols, ASCII symbols, and punctuation marks. Furthermore, the text characteristics were handled using different methods such as tokenization, case folding, non-standard word handling, stemming, and stop-word removal. The results were later forwarded to the feature extraction stage to produce a scoring matrix by combining the TF-IDF and lexicon methods.

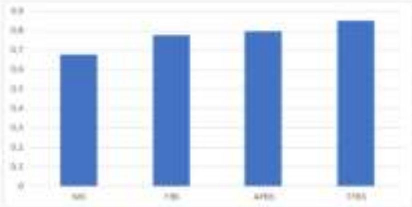


**Fig. 5.** The accuracy values obtained from using different feature extraction methods

Figure 5 shows the accuracy values obtained from using different feature extraction methods. It is important to note that the focus was placed on the tweet classification using five classes of fear, angry, love, sad, and happy. The results showed that the proposed TFBS method had a higher accuracy value of 0.85 compared to other methods.

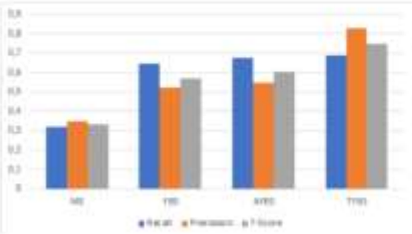


**Fig. 6.** Result of the Recall, Precision, and F-Score measurements for each method

As shown in Figure 6, the performance was also evaluated by calculating the precision, recall, and F-score values for each extraction method used, and the proposed TFBS method was observed to have the highest values while ME had the lowest.

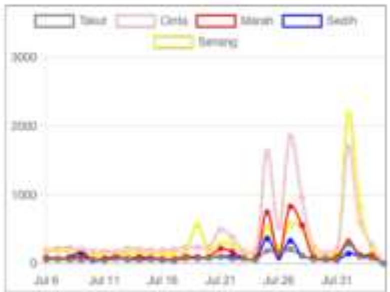



**Fig. 7.** Trend of public reactions in 06 July 2022 to 04 Aug 2022

The next experiment was to classify public reactions during the COVID-19 pandemic using the TFBS method with a focus on happy, sad, fear, angry, and love categories. It was discovered from the crawling analysis in Figure 7 that public reactions associated with "love" were higher than others. On July 24-31, 2022 the reactions to the pandemic were observed to be normal with the assumption that it could not attack Indonesia because of the ongoing vaccination program and the creation of herd immunity. The people thought those that were not disciplined with the implementation of the health protocols and refused vaccination were most susceptible to the infection. The "angry" class was found in second place on July 25-27 2022 due to the increase in COVID-19 cases in some areas that have returned to normal activities including schools that have been closed to implement online learning. It was also discovered that other viruses such as monkeypox were becoming an issue in these communities.


**Fig. 8.** Real- time on the dashboard of the monitoring system every 2 minutes in the last 1 hour

Moreover, in Figure 8, the trend of public reactions was also visualized in real- time on the dashboard of the monitoring system with a focus on every 2 minutes in the last 1 hour. It was discovered that the trend from 05.20 AM to approximately 06.00 tends more toward fear reactions. This simply indicates the general feeling of the people in the community regarding COVID-19.

The classifications are indicated by visualization in the form of a word cloud as shown in Figures 9 to 13. It was discovered that the words often observed in each class vary. For example, the frequent words in "happy" class include health, protocol, and discipline, those in "sad" are infected, sick, and fever, "love" had health, protocol, and discipline, "fear" recorded results, examination, and ministry of health while "angry" had infected, vaccinated, and positive.

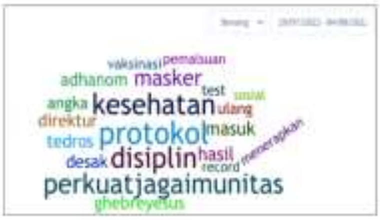



**Fig. 9.** Popular topic of Word Cloud that associated with happy reaction

Figure 9 shows the trend of public reactions in happy category through the word cloud consisting of several frequent words that are widely discussed in the community and formed the opinion of the people. Some of these include protocol, health, masks, and discipline that were trending because, at the time, the enforcement of health protocol discipline was the main thing and had become an issue of public awareness.


**Fig. 10.** Popular topic of Word Cloud that associated with sad reaction

Figure 10 indicates the public reactions for the sad category are dominated by frequent words such as infected, sick, fever, positive, and cough. They were appearing often at the time because work and school activities of people were conducted offline. This led to several positive cases and other symptoms of COVID-19.

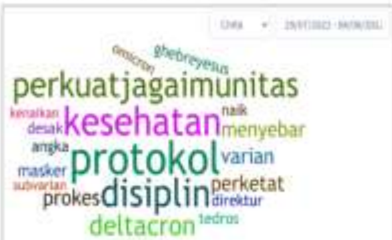



**Fig. 11.** Popular topic of Word Cloud that associated with love reaction

Figure 11 shows that the frequent words in the "love" category include health, protocol, discipline, progress, increase, deltacron, variant, and mask. These are almost the same as those observed in the happy category but the love aspect focuses on the hope for the better. During this period, new variants of COVID-19 began to appear such as omicron and deltacron but the public expresses their love while maintaining health protocols and discipline through the usage of masks.


**Fig. 12.** Popular topic of Word Cloud that associated with fear reaction

Figure 12 shows that the words trending in the fear category are dominated by the ministry of health, health, laboratory, examination, discipline, and protocol. The focus was placed more on the issue of an increase in COVID-19 patients, the emergence of new variants and the monkeypox virus during the period, and the need for the Ministry of Health to respond through appropriate policies. The public was always vigilant, afraid, and kept reminding each other to always be disciplined in maintaining health protocols.


**Fig. 13.** Popular topic of Word Cloud that associated with angry reaction

The trending words related to the fear category were found in Figure 13 to be dominated by words such as sick, affected, positive, death, and vaccine. The cause of these reactions was observed to be almost the same as the sad category.

**Fig. 14.** Visualization of the COVID-19 public reaction monitoring system report

Some words in the word cloud from each category that are believed to have appeared due to their high frequency of occurrence in text tweets written by users. This means they reflect the opinion developed in the community. Figure 14 shows a visualization of the report from the COVID-19 public reaction monitoring system with a focus on the tweet recapitulation of each reaction category. It also indicated the graph of people's reactions in real-time which can also be determined at a certain time.

## 5. Conclusion

This research developed a sentiment analysis model using the TF-IDF and lexical methods to identify public reactions to COVID- 19. The prediction accuracy measurement showed that the proposed method was more accurate than the others due to its value of 0.854. It was also discovered from the Recall, Precision, and F-Score evaluation that the ME method had poor performance while the proposed TFBS had the best followed by AFBS and FBS. The TFBS also ranked higher compared to the other methods because it correctly identified and classified tweets. Moreover, the experiments conducted to classify the tendency of public reactions in dealing with the COVID-19 pandemic using the method showed the possibility of identifying and visualizing several categories on the dashboard of the community reaction monitoring system.

## References


[1] R. P. Pratama and A. Tjahyanto, "The influence of fake accounts on sentiment analysis related to COVID-19 in Indonesia," in Procedia Computer Science, 2021, vol. 197, no. 2021, pp. 143–150, doi: 10.1016/j.procs.2021.12.128.

[2] A. D. Fakhira, A. S. Pawitra, K. C. Diyanah, E. A. Wikurendra, I. Nagy, and N. S. M. Abdeljawad, "Awareness of Doing 3M (Wearing Mask, Physical Distancing, Washing Hands) During Pandemic Era in Rural and Urban Families," J. Kesehat. Lingkung., vol. 13, no. 2, p. 94, 2021, doi: 10.20473/jkl.v13i2.2021.94-101.

[3] A. K. and M.-V. Dorothée Allain-Dupré, Isabelle Chatry and Michalun, "The territorial impact of COVID-19 :

Managing The Crisis Across Levels of Government," 2020. [Online]. Available: https://www.oecd.org/coronavirus/policy-responses/the-territorial-impact-of-covid-19-managing-the-crisis-across-levels-of-government-d3e314e1/.

[4] L. Nahar, Z. Sultana, N. Iqbal, and A. Chowdhury, "Sentiment Analysis and Emotion Extraction: A Review of Research Paradigm," 1st Int. Conf. Adv. Sci. Eng. Robot. Technol. 2019, ICASERT 2019, no. December, 2019, doi: 10.1109/ICASERT.2019.8934654.

[5] Z. Drus and H. Khalid, "Sentiment analysis in social media and its application: Systematic literature review," Procedia Comput. Sci., vol. 161, pp. 707–714, 2019, doi: 10.1016/j.procs.2019.11.174.

[6] A. Pertiwi, A. Triayudi, and E. T. E. Handayani, "Sentiment Analysis of the Impact of Covid-19 on Indonesia's Economy through Social Media Using the ANN Method," J. Mantik, vol. 4, no. 1, pp. 605–612, 2020, [Online]. Available: https://iocscience.org/ejournal/index.php/mantik.

[7] P. Yadav and D. Pandya, "SentiReview: Sentiment Analysis based on Text and Emoticons," in International Conference on Innovative Mechanisms for Industry Applications, 2017, pp. 467–472.

[8] S. Peng, G. Wang, and D. Xie, "Social Influence Analysis in Social Networking Big Data: Opportunities and Challenges," IEEE Netw., vol. 31, no. 1, pp. 11–17, 2017, doi: 10.1109/MNET.2016.1500104NM.

[9] A. Go, L. Huang, and R. Bhayani, "Twitter Sentiment Analysis," 2009.

[10] C. R. Machuca, C. Gallardo, and R. M. Toasa, "Twitter sentiment analysis on coronavirus: Machine learning approach," J. Phys. Conf. Ser., vol. 1828, no. 1, 2021, doi: 10.1088/1742-6596/1828/1/012104.

[11] N. N. Alabid and Z. D. Katheeth, "Sentiment analysis of twitter posts related to the covid-19 vaccines," Indones. J. Electr. Eng. Comput. Sci., vol. 24, no. 3, pp. 1727–1734, 2021, doi: 10.11591/ijeecs.v24.i3.pp1727-1734.

[12] J. Srikanth, A. Damodaram, Y. Teekaraman, R. Kuppusamy, and A. R. Thelkar, "Sentiment Analysis on COVID-19 Twitter Data Streams Using Deep Belief Neural Networks," Comput. Intell. Neurosci., 2022, doi: https://doi.org/10.1155/2022/8898100.

[13] D. A. Efrilianda, E. N. Dianti, and O. G. Khoirunnisa, "Analysis of twitter sentiment in COVID-19 era using fuzzy logic method," J. Soft Comput. Explor., vol. 1, no. 1, pp. 1–5, 2021, doi: 10.52465/joscex.v2i1.12.

[14] T. N. Simanjuntak and S. Pramana, "Sentiment Analysis on Overseas Tweets on the Impact of COVID-19 in Indonesia," Indones. J. Stat. Its Appl., vol. 5, no. 2, pp. 304–313, 2021, doi: 10.29244/ijsa.v5i2p304-313.

[15] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," Artif. Intell. Rev., 2022, doi: https://doi.org/10.1007/s10462-022-10144-1.

[16] T. N. Prakash and A. Aloysius, "Data Preprocessing in Sentiment Analysis using Twitter Data," Int. Educ. Appl. Res. J., vol. 3, no. 7, 2019.

[17] A. F. Hidayatullah and M. R. Ma'arif, "Pre-processing Tasks in Indonesian Twitter Messages," J. Phys. Conf. Ser. Vol. 801, Int. Conf. Comput. Appl. Informatics - IOP Publ. Ltd, 2016, doi: 10.1088/1742-6596/801/1/012072.

[18] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," Procedia Comput. Sci., vol. 152, pp. 341–348, 2019, doi: 10.1016/j.procs.2019.05.008.

[19] Kusrini and M. Mashuri, "Sentiment Analysis In Twitter Using Lexicon Based and Polarity Multiplication," 2019, doi: 10.1109/ICAIIT.2019.8834477.

[20] [N. Medagoda, S. Shanmuganathan, and J. L. Whalley, "Sentiment Lexicon Construction," 2015, [Online]. Available: 10.1109/ICNC.2015.7378094.

| | |
|---|---|
| Seller Username | researchpublicationat |
| Seller Email Address | editor@thedesignengineering.com |

### Purchase Details

| | |
|---|---|
| Payment ID | MOJO2c15J05A65174994 |
| Buyer Name | Djoko Adi Widodo |
| Buyer Email Address | djokoadiwidodo19@gmail.com |
| Buyer Phone Number | +628122541733 |
| Purchase Date | Dec. 15, 2022 |
| Price | INR. 37000.00 |
| Total Purchase Amount | INR. 37000.00 |

Powered by instamojo

## COPY RIGHT DECLARATION

### International Journal of Intelligent Systems and Applications in Engineering (IJISAE)

**Title of the Article:** Sentiment Analysis of Twitter Media for Public Reaction Identification on COVID-19 Monitoring System using Hybrid Feature Extraction Method

**Authors**

I/We hereby assign copyright of the article named above (the Work), to the publisher, International Journal of Intelligent Systems and Applications in Engineering (IJISAE). I/We understand that International Journal of Intelligent Systems and Applications in Engineering (IJISAE) will act on my/our behalf to publish, reproduce, distribute and transmit the Work and will authorize other reputable third parties (such as document delivery services) to do the same. I/We warrant that the Work has not been published before in its current or a substantially similar form and is not under consideration for another publication, does not contain any unlawful statements. I/We warrant that "proof of consent" has been obtained for studies of named organizations and people. All authors have received a final version of the Work, take responsibilityfor the content and agree to its submission. I/We assert my/our moral rights to be identified as the author/s of the Work.

**Note:** In case if all authors are not available then only one author's signature is sufficient.

1. Signature ____________________

Author Name Djoko Adi Widodo Date December 12nd, 2022 Place Semarang, Indonesia

2. Signature ____________________

Author Name Nur Iksan Date December 12nd, 2022 Place Semarang, Indonesia

3. Signature ____________________

Author Name Budi Sunarko Date December 12nd, 2022 Place Semarang, Indonesia

4. Signature ____________________

Author Name Date Place

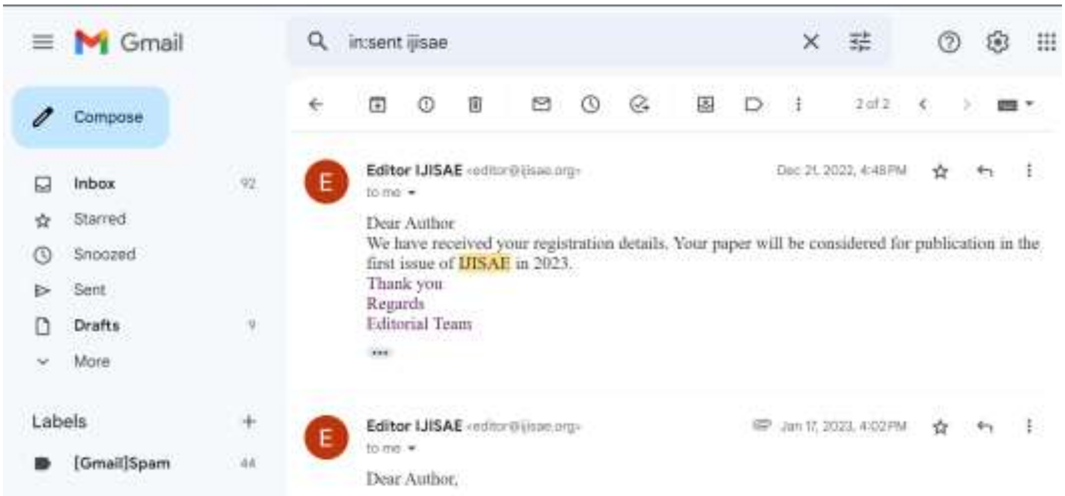
5. Signature ____________________

Author Name Date Place

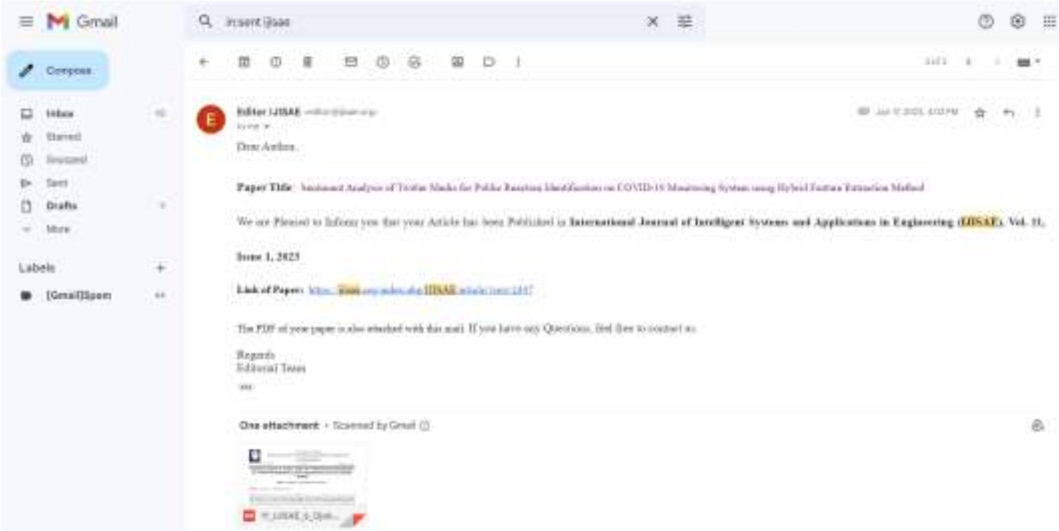
6. Signature ____________________

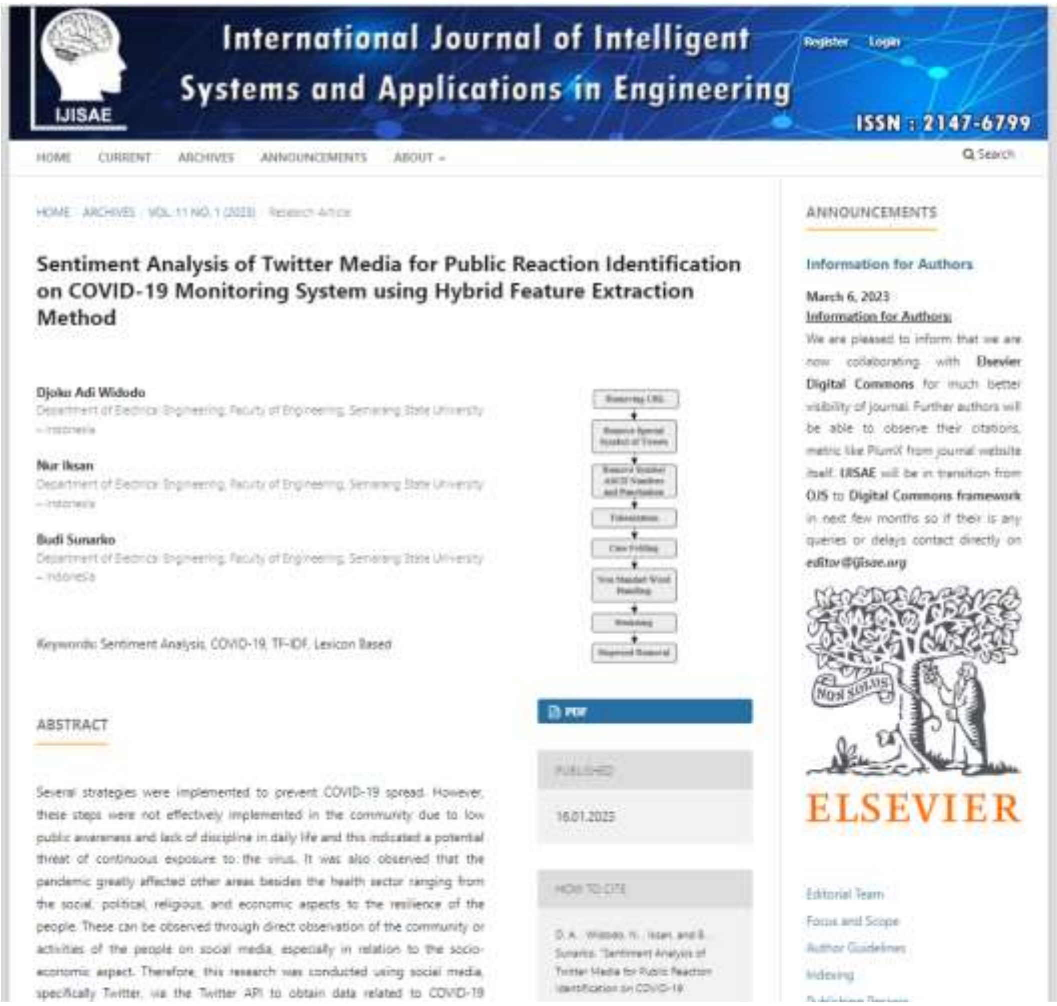
Author Name Date Place

## 11. Pemberitahuan registrasi untuk publikasi diterima



## 12. Pemberitahuan bahwa manuskrip telah dipublikasi online.



International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

# Sentiment Analysis of Twitter Media for Public Reaction Identification on COVID-19 Monitoring System using Hybrid Feature Extraction Method

**Djoko Adi Widodo*[1], Nur Iksan[1], Budi Sunarko[1]**

**Submitted:** 19/10/2022 **Accepted:** 31/12/2022

**Abstract:** Several strategies were implemented to prevent COVID-19 spread. However, these steps were not effectively implemented in the community due to low public awareness and lack of discipline in daily life and this indicated a potential threat of continuous exposure to the virus. It was also observed that the pandemic greatly affected other areas besides the health sector ranging from the social, political, religious, and economic aspects to the resilience of the people. These can be observed through direct observation of the community or activities of the people on social media, especially in relation to the socio-economic aspect. Therefore, this research was conducted using social media, specifically Twitter, via the Twitter API to obtain data related to COVID-19 pandemic in Indonesia. In this research, a sentiment analysis method was developed in this research to identify public opinion related to the spread of the COVID-19 virus and its social impact on society. This was achieved using the TF-IDF and Lexical methods for feature extraction and Naive Bayes for classification. This research used a dataset obtained through Twitter using the keyword "COVID-19" in Indonesian and manually labelled using 5 categories of reactions i.e., fear, angry, love, sad, and happy. The prediction accuracy values showed that the proposed TFBS method had a higher accuracy value of 0.85 compared to other methods. The performance was also evaluated by calculating the precision, recall, and F-score values for each extraction method used, and the proposed TFBS method was observed to have the highest values while ME had the lowest.

***Keywords:*** *Sentiment Analysis, COVID-19, TF-IDF, Lexicon Based*

## 1. Introduction

World Health Organization (WHO) declared the COVID-19 virus a pandemic due to its spread to all countries in the world, thereby, becoming a serious concern for the public, government, and world health institutions [1]. The virus started emerging in Indonesia in early January 2020 and the data from the handling team showed an increasing trend with nearly 600,000 people reported to be exposed and approximately 18,000 died. Several strategies were implemented to prevent further spread such as washing hands, wearing masks, quarantine, social distancing, working from home, and others [2]. However, these steps were not effectively implemented in the community due to low public awareness and lack of discipline in daily life and this indicated a potential threat of continuous exposure to the virus. This led the government to introduce another strategy which involved launching a vaccination program for all elements of society to prevent infection.

This vaccination program was one of the effective efforts made to suppress the spread with the main goal of providing immunity against COVID-19 virus infection and ensuring the formation of herd or group immunity. Furthermore, the spread was also prevented by providing information to increase public awareness on the best way to conduct their activities during the period.

In addition to the health sector, the pandemic also significantly affected the economic, religious, security, resilience, political, and social sectors [3]. These were not identified directly as the case with the health sector but through the observation of community activities using different media such as Instagram, Facebook, Twitter, and others. These media serve as social sensors for communities through the uploading of ideas, thoughts, and opinions in the forms of text, video, photos, and audio. The contents are focused on activities and conditions of the things the people feel and see in their surrounding environment which are subsequently sha red on social media platforms. Meanwhile, sentiment analysis can be used to identify and predict future conditions using the information [4–6]. A sentiment analysis method was developed in this research to identify public opinion related to the spread of the COVID-19 virus and its social impact on society [7–9]. This was achieved using the TF-IDF and Lexical methods for feature extraction and Naive Bayes for classification.

[1] *Department of Electrical Engineering, Faculty of Engineering, Semarang State University – Indonesia*
[2] *Department of Electrical Engineering, Faculty of Engineering, Semarang State University – Indonesia*
[3] *Department of Electrical Engineering, Faculty of Engineering, Semarang State University – Indonesia*
* *Corresponding Author Email: djokoadiwidodo@mail.unnes.ac.id*

## 2. Related Works

Sentiment analysis has become an interesting topic and its methods have been widely developed, especially on Twitter concerning the COVID-19 spread. It was used by Machuca [10] to determine negative and positive sentiments through Machine Learning (ML) and Natural Language Processing (NLP) methods. The Logistic Regression Algorithm was used as an ML algorithm to classify feelings or emotions regarding the COVID-19 topic on Twitter in 2020. Meanwhile, the TF-IDF method was used to extract NLP features and the classification accuracy value was found to be 78.5%. Alabid & Katheeth [11] also developed a model framework to classify public sentiment and opinion regarding the implementation of the COVID-19 vaccination program. The research further compared 2 ML methods including Support Vector Machine (SVM) and Naive Bayes (NB) to classify sentiment polarity into positive, negative, and neutral. The experimental results showed that the NB method had a higher accuracy performance than SVM with 0.81 and 0.75 respectively. Srikanth et al., [12] also conducted a sentiment analysis using a combination of preprocessing methods and word embedding to perform feature extraction. The Deep Belief Neural Network (DBN) algorithm was used to classify tweets through a pseudo-labeling process and the experimental results showed that the DBN algorithm achieved an accuracy value of 90.3% which is better than the other methods. Moreover, Efrilianda [13] developed a sentiment analysis method to determine positive, negative, and neutral polarity using the Backpropagation Multi- Layer Perceptron (MLP) algorithm and the results showed an accuracy value of 70%. It was observed that these studies did not consider the development of a feature extraction method as an important factor in increasing the accuracy of sentiment classification. They also only classified sentiments into negative, positive, and neutral polarities instead of emotions and feelings. Therefore, this research developed a feature extraction method by combining TF-IDF with the lexical method. Sentiment analysis was also developed concerning the community's reactions to the COVID-19 pandemic with a focus on happy, sad, angry, fear, and love categories.

## 3. Method

Sentiment analysis is also known as opinion mining and it is normally used to process natural discussion or text commonly referred to as NLP to determine the sentiments embedded in an opinion. It usually categorizes the opinion sentences in the form of polarity which include Negative, Positive, and Neutral [9, 14]. It can also be used for classification in other forms such as happy, sad, angry, fear, and love. The steps in the sentiment analysis process are shown in the following Figure 1:

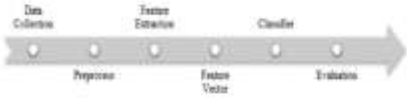


**Fig. 1**. The steps in the sentiment analysis process

The sentiment analysis consists of 6 stages of data collection, data pre-processing, feature extraction, vectorization, classification, and evaluation. The data collection stage was used to retrieve text data from Twitter through the API [15]. The data pre-processing was used to prepare the collected data for extraction into features [16] and this is important to determine the performance of the sentiment analysis process later. Overall, this stage is divided into two which include general pre-processing associated with the text contained in the tweet and special pre-processing related to the characteristics of the tweet text [17]. Figure 2 shows the stages of data pre-processing conducted in this research.


**Fig. 2.** Pre-processing stages

1. Removal URL by filtering the tweet text to remove URLs such as "HTTP://".
2. Removal of the special symbol in the tweets such as hashtags (#), retweets (RT), and usernames (@).
3. Tokenization such that the tweet text in the form of a sentence was be divided into several constituent words or phrases called tokens using the tokenize() function. The function was obtained from the RegexpTokenizer() package in the NLTK library.
4. Case folding which involved changing some words from the tokenization process to lowercase letters

using the lower() function.

5. Nonstandard word handling which involved returning the detected unstandardized words such as errors in the use of spelling, abbreviations, slang words, and elongated words to standard form.

6. The stemming process plays an important role in finding the root word of a derivative word. This process will remove affixes to a word. The method used in this process will be different in each language. In Indonesian usage, the stemming process uses the Literature library, while in English it uses the NLTK library.

7. Stopword Removal stage which involved removing common words that do not have a significant effect on the tweet sentence. This was achieved through the Literature and NLTK libraries.

The next stage is extraction which is very fundamental in the sentiment analysis process [18]. It involves converting the pre- processed text into vector form and was conducted in this research using the TF-IDF approach which normally extracts sentence features based on the frequency of word occurrence. Moreover, the frequency term was calculated by determining a feature based on the consistent appearance of a word. The features represent a document and have a high weight. In contrast to TF-IDF, the frequency of a word occurrence in a document is usually compared with an entire database. In a situation the word appears too often in several documents, it cannot be used as a feature because it is considered a general word that does not represent a specific document. The TF-IDF weight was calculated using the following Equations (1) and (2):

$$tf.idf_{t,d} = tf_{td} \times idf_t \tag{1}$$

$$idf_t = \log \frac{N}{df_t} \tag{2}$$

The lexical method was also applied in this research and it involves using the word meaning to determine the sentence features. The accuracy of the sentiment analysis is higher when the word meaning is more complete [19, 20]. The use of this method usually focuses on applying word lists such as "clean", "healthy", "good", and "safe" to determine positive polarity and others such as "sloppy", "bad", "and slow" to determine the negative polarity. This feature extraction for the English lexicon was conducted using the library from the Google Translate API and the lexicon was translated first after which the sentiment scores were calculated from sentiwordnet using Equations (3) and (4).

$$Score = \sum \frac{(PosScore - NegScore)}{tot_{index}} \tag{3}$$

$$Sentence_{Score} = \sum Score \tag{4}$$

Meanwhile, the feature extraction of the Indonesian language lexicon was conducted by calculating the appearance of the word opinion in each sentence after which the positive and negative sentiments were determined using Equations (5) and (6).

$$Score = \sum frek_{opini_{word}} \tag{5}$$

$$Sentence_{Score} = \sum_{i=0}^{n} \sum Score_i \tag{6}$$

This means TF-IDF and Lexical methods which are statistical and semantic approaches respectively were combined in this research to determine the weighting of the feature extraction process. The proposed sentiment analysis model is presented in the following Figure 3.

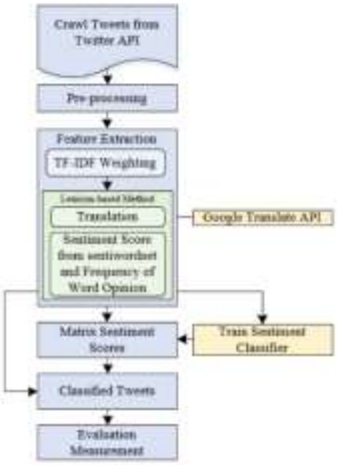



**Fig. 3.** Proposed Sentiment Analysis Model

At the classification stage, the Naive Bayes algorithm was used to determine the tweets consisting of 5 classes which include fear, angry, love, sad, and happy. Furthermore, the performance evaluation was conducted by comparing the Feature-Based Sentiment algorithm using TF-IDF and Lexicon (TFBS) with several other methods such as Maximum Entropy (ME), Feature-Based Sentiment (FBS), and Augmented Feature-Based Sentiment (AFBS).

Performance evaluation was conducted by calculating the accuracy, precision, recall, and F-Score. The accuracy value was calculated by comparing the number of correct data and the total amount of data. A higher accuracy value indicates a better method. Moreover, precision, recall, and F-score were measured to determine any deviations in the data. Precision value was determined by comparing the number of relevant classification data with the total amount of data from a particular class. The recall value was obtained from the comparison of the amount of relevant data with the total relevant data while the F-Score value was determined by calculating the average value of precision and recall. The equations used to calculate precision, recall, and f-score values are stated as Equations (7), (8) and (9).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (7)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (8)$$

$$F - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (9)$$

## 4. Result and Discussion

This research used a dataset obtained through Twitter using the keyword "COVID-19" in Indonesian. The dataset was retrieved through a crawling process using the Tweepy library in Python for as many as 3148 lines. It was subsequently passed through the normalization stage to eliminate duplicated data such as the tweets reposted by other users or retweets (RT). The dataset was manually labelled using 5 categories of reactions as shown in Table 1.

**Table 1.** Labelled Dataset using 5 categories of reactions

| Text Tweet (Original) | Text Tweet (after translated in English) | Labels |
|---|---|---|
| Positif covid-19 bukan kutukan dan bukan aib. Terima kasih sudah berbagi dan tetap semangat.. | Positive covid-19 is not a curse and not a disgrace. Thank you for sharing and keep the spirit. | Happy |
| sedih banget kalo denger cerita perawat yang jadi satgas covid-19. banyak pasien covid yang bandel dan susah dibilangin ternyata | It's really sad to hear the story of the nurse who became the Covid-19 task force. many covid patients are stubborn and hard to say it turns out | Sad |
| gak respect banget.. ini ada sodara meninggal di RS padahal karena riwayat stroke, udah dimandiin dan diangkat pas beres malah keluar surat positif COVID-19?! Sumpah baru pertama kali ngeliat langsung kasus orang meninggal dicovidkan dan bener adanya | really no respect.. this is a friend who died in the hospital even though because of a history of stroke, he was washed and picked up when it was done, instead a positive letter for COVID-19 came out?! I swear it's the first time I've seen a case of someone who died being COVID-19 and it's true | Angry |
| Besok mau di vaksin covid :( Hrusnya seneng ga sih Kan nntinya dpt perlindungan, jujur malah takut :( | Tomorrow I'm going to be vaccinated against covid :( Should I be happy or not, I'm going to get protection, to be honest I'm scared :( | Fear |
| Rindu suasana belajar mengajar sebelum covid mewabah di negri ini | Missing the teaching and learning atmosphere before the covid outbreak in this country | Love |


**Fig. 4.** Example of a text tweet on Twitter

The tweets contain some useless information or noise such as hashtags (#), Html (http://www), mentions (@), numbers, local language usage, and abbreviations. The initial stage in the sentiment analysis process is the pre-processing which involves removing the URL, special symbol, punctuation, tokenization, case folding, and stemming as shown in Figure 2. The useless information was removed by omitting URLs, special symbols, ASCII numbers, and punctuations. Figure 4 shows an example of a tweet containing noise and this means there is a need for pre- processing. The tweet text in Figure 4 was pre-processed and the results are presented in the following Table 2.

**Table 2.** Cleansing at the Pre-processing Stage

| Preprocessing stage | Original Tweet | Tweet in English |
|---|---|---|
| Tweet | " Orang-orang begitu mana paham covid-19, dijelasin berulangpun tetep ngeyel. What should i do? . . . Yaudah biarin aja,kalo positif apalagi sampe bergejala ga akan diurus. Simpan tenagamu untuk yang benar-benar membutuhkan ya nakes 😢. Kita kuat kita bisa survive" | " How do people understand covid-19, even when they explain it again and again, it's still annoying. What should I do? . . . Well, just let it be, if it's positive, especially if you have symptoms, it won't be taken care of. Save your energy for those who really need it, guys. We are strong we can survive" |
| Preprocessing Results | "orang paham covid jelas ulang udah biar saja positif gejala urus simpan tenaga butuh ya nakes kuat survive" | "people who understand that covid is being explained again, they keep yelling, what should i do, let's just leave it if it's positive until the symptoms don't take care of saving energy, we need strong health workers to survive" |
| Tokenization Result | [orang] [paham] [covid] [jelas] [ulang] [udah] [biar] [saja] [positif] [gejala] [urus] [simpan] [tenaga] [butuh] [nakes] [kuat] [survive] | [people] [understand] [covid] [clear] [repeat] [already] [let it be] [just] [positive] [symptoms] [manage] [save] [energy] [need] [nakes] [strong] [survival] |

Several processes were implemented at the pre-processing stage and these include the deletion of emoticons, URLs, special symbols, ASCII symbols, and punctuation marks. Furthermore, the text characteristics were handled using different methods such as tokenization, case folding, non-standard word handling, stemming, and stop-word removal. The results were later forwarded to the feature extraction stage to produce a scoring matrix by combining the TF-IDF and lexicon methods.

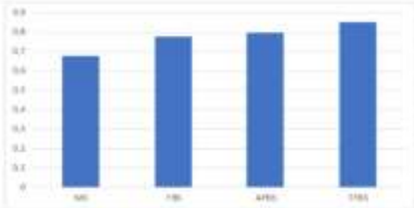


**Fig. 5.** The accuracy values obtained from using different feature extraction methods

Figure 5 shows the accuracy values obtained from using different feature extraction methods. It is important to note that the focus was placed on the tweet classification using five classes of fear, angry, love, sad, and happy. The results showed that the proposed TFBS method had a higher accuracy value of 0.85 compared to other methods.

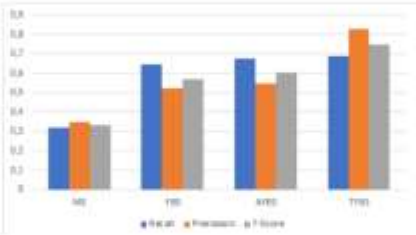


**Fig. 6.** Result of the Recall, Precision, and F-Score measurements for each method

As shown in Figure 6, the performance was also evaluated by calculating the precision, recall, and F-score values for each extraction method used, and the proposed TFBS method was observed to have the highest values while ME had the lowest.

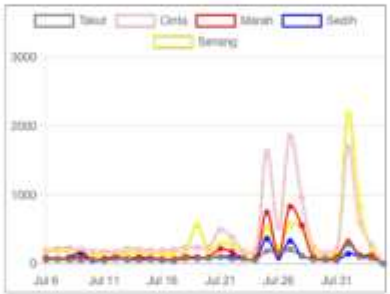



**Fig. 7.** Trend of public reactions in 06 July 2022 to 04 Aug 2022

The next experiment was to classify public reactions during the COVID-19 pandemic using the TFBS method with a focus on happy, sad, fear, angry, and love categories. It was discovered from the crawling analysis in Figure 7 that public reactions associated with "love" were higher than others. On July 24-31, 2022 the reactions to the pandemic were observed to be normal with the assumption that it could not attack Indonesia because of the ongoing vaccination program and the creation of herd immunity. The people thought those that were not disciplined with the implementation of the health protocols and refused vaccination were most susceptible to the infection. The "angry" class was found in second place on July 25-27 2022 due to the increase in COVID-19 cases in some areas that have returned to normal activities including schools that have been closed to implement online learning. It was also discovered that other viruses such as monkeypox were becoming an issue in these communities.


**Fig. 8.** Real- time on the dashboard of the monitoring system every 2 minutes in the last 1 hour

Moreover, in Figure 8, the trend of public reactions was also visualized in real- time on the dashboard of the monitoring system with a focus on every 2 minutes in the last 1 hour. It was discovered that the trend from 05.20 AM to approximately 06.00 tends more toward fear reactions. This simply indicates the general feeling of the people in the community regarding COVID-19.

The classifications are indicated by visualization in the form of a word cloud as shown in Figures 9 to 13. It was discovered that the words often observed in each class vary. For example, the frequent words in "happy" class include health, protocol, and discipline, those in "sad" are infected, sick, and fever, "love" had health, protocol, and discipline, "fear" recorded results, examination, and ministry of health while "angry" had infected, vaccinated, and positive.

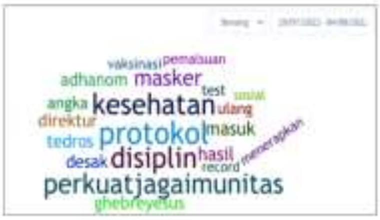



**Fig. 9.** Popular topic of Word Cloud that associated with happy reaction

Figure 9 shows the trend of public reactions in happy category through the word cloud consisting of several frequent words that are widely discussed in the community and formed the opinion of the people. Some of these include protocol, health, masks, and discipline that were trending because, at the time, the enforcement of health protocol discipline was the main thing and had become an issue of public awareness.


**Fig. 10.** Popular topic of Word Cloud that associated with sad reaction

Figure 10 indicates the public reactions for the sad category are dominated by frequent words such as infected, sick, fever, positive, and cough. They were appearing often at the time because work and school activities of people were conducted offline. This led to several positive cases and other symptoms of COVID-19.


**Fig. 11.** Popular topic of Word Cloud that associated with love reaction

Figure 11 shows that the frequent words in the "love" category include health, protocol, discipline, progress, increase, deltacron, variant, and mask. These are almost the same as those observed in the happy category but the love aspect focuses on the hope for the better. During this period, new variants of COVID-19 began to appear such as omicron and deltacron but the public expresses their love while maintaining health protocols and discipline through the usage of masks.


**Fig. 12.** Popular topic of Word Cloud that associated with fear reaction

Figure 12 shows that the words trending in the fear category are dominated by the ministry of health, health, laboratory, examination, discipline, and protocol. The focus was placed more on the issue of an increase in COVID-19 patients, the emergence of new variants and the monkeypox virus during the period, and the need for the Ministry of Health to respond through appropriate policies. The public was always vigilant, afraid, and kept reminding each other to always be disciplined in maintaining health protocols.


**Fig. 13.** Popular topic of Word Cloud that associated with angry reaction

The trending words related to the fear category were found in Figure 13 to be dominated by words such as sick, affected, positive, death, and vaccine. The cause of these reactions was observed to be almost the same as the sad category.


**Fig. 14.** Visualization of the COVID-19 public reaction monitoring system report

Some words in the word cloud from each category that are believed to have appeared due to their high frequency of occurrence in text tweets written by users. This means they reflect the opinion developed in the community. Figure 14 shows a visualization of the report from the COVID-19 public reaction monitoring system with a focus on the tweet recapitulation of each reaction category. It also indicated the graph of people's reactions in real-time which can also be determined at a certain time.

## 5. Conclusion

This research developed a sentiment analysis model using the TF-IDF and lexical methods to identify public reactions to COVID- 19. The prediction accuracy measurement showed that the proposed method was more accurate than the others due to its value of 0.854. It was also discovered from the Recall, Precision, and F-Score evaluation that the ME method had poor performance while the proposed TFBS had the best followed by AFBS and FBS. The TFBS also ranked higher compared to the other methods because it correctly identified and classified tweets. Moreover, the experiments conducted to classify the tendency of public reactions in dealing with the COVID-19 pandemic using the method showed the possibility of identifying and visualizing several categories on the dashboard of the community reaction monitoring system.

## References


[1] R. P. Pratama and A. Tjahyanto, "The influence of fake accounts on sentiment analysis related to COVID-19 in Indonesia," in Procedia Computer Science, 2021, vol. 197, no. 2021, pp. 143–150, doi: 10.1016/j.procs.2021.12.128.

[2] A. D. Fakhira, A. S. Pawitra, K. C. Diyanah, E. A. Wikurendra, I. Nagy, and N. S. M. Abdeljawad, "Awareness of Doing 3M (Wearing Mask, Physical Distancing, Washing Hands) During Pandemic Era in Rural and Urban Families," J. Kesehat. Lingkung., vol. 13, no. 2, p. 94, 2021, doi: 10.20473/jkl.v13i2.2021.94-101.

[3] A. K. and M.-V. Dorothée Allain-Dupré, Isabelle Chatry and Michalun, "The territorial impact of COVID-19:

Managing The Crisis Across Levels of Government," 2020. [Online]. Available: https://www.oecd.org/coronavirus/policy-responses/the-territorial-impact-of-covid-19-managing-the-crisis-across-levels-of-government-d3e314e1/.

[4] L. Nahar, Z. Sultana, N. Iqbal, and A. Chowdhury, "Sentiment Analysis and Emotion Extraction: A Review of Research Paradigm," 1st Int. Conf. Adv. Sci. Eng. Robot. Technol. 2019, ICASERT 2019, no. December, 2019, doi: 10.1109/ICASERT.2019.8934654.

[5] Z. Drus and H. Khalid, "Sentiment analysis in social media and its application: Systematic literature review," Procedia Comput. Sci., vol. 161, pp. 707–714, 2019, doi: 10.1016/j.procs.2019.11.174.

[6] A. Pertiwi, A. Triayudi, and E. T. E. Handayani, "Sentiment Analysis of the Impact of Covid-19 on Indonesia's Economy through Social Media Using the ANN Method," J. Mantik, vol. 4, no. 1, pp. 605–612, 2020, [Online]. Available: https://iocscience.org/ejournal/index.php/mantik.

[7] P. Yadav and D. Pandya, "SentiReview: Sentiment Analysis based on Text and Emoticons," in International Conference on Innovative Mechanisms for Industry Applications, 2017, pp. 467–472.

[8] S. Peng, G. Wang, and D. Xie, "Social Influence Analysis in Social Networking Big Data: Opportunities and Challenges," IEEE Netw., vol. 31, no. 1, pp. 11–17, 2017, doi: 10.1109/MNET.2016.1500104NM.

[9] A. Go, L. Huang, and R. Bhayani, "Twitter Sentiment Analysis." 2009.

[10] C. R. Machuca, C. Gallardo, and R. M. Toasa, "Twitter sentiment analysis on coronavirus: Machine learning approach," J. Phys. Conf. Ser., vol. 1828, no. 1, 2021, doi: 10.1088/1742-6596/1828/1/012104.

[11] N. N. Alabid and Z. D. Katheeth, "Sentiment analysis of twitter posts related to the covid-19 vaccines," Indones. J. Electr. Eng. Comput. Sci., vol. 24, no. 3, pp. 1727–1734, 2021, doi: 10.11591/ijeecs.v24.i3.pp1727-1734.

[12] J. Srikanth, A. Damodaram, Y. Teekaraman, R. Kuppusamy, and A. R. Thelkar, "Sentiment Analysis on COVID-19 Twitter Data Streams Using Deep Belief Neural Networks," Comput. Intell. Neurosci., 2022, doi: https://doi.org/10.1155/2022/8898100.

[13] D. A. Efrilianda, E. N. Dianti, and O. G. Khoirunnisa, "Analysis of twitter sentiment in COVID-19 era using fuzzy logic method," J. Soft Comput. Explor., vol. 1, no. 1, pp. 1–5, 2021, doi: 10.52465/joscex.v2i1.12.

[14] T. N. Simanjuntak and S. Pramana, "Sentiment Analysis on Overseas Tweets on the Impact of COVID-19 in Indonesia," Indones. J. Stat. Its Appl., vol. 5, no. 2, pp. 304–313, 2021, doi: 10.29244/ijsa.v5i2p304-313.

[15] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," Artif. Intell. Rev., 2022, doi: https://doi.org/10.1007/s10462-022-10144-1.

[16] T. N. Prakash and A. Aloysius, "Data Preprocessing in Sentiment Analysis using Twitter Data," Int. Educ. Appl. Res. J., vol. 3, no. 7, 2019.

[17] A. F. Hidayatullah and M. R. Ma'arif, "Pre-processing Tasks in Indonesian Twitter Messages," J. Phys. Conf. Ser. Vol. 801, Int. Conf. Comput. Appl. Informatics - IOP Publ. Ltd, 2016, doi: 10.1088/1742-6596/801/1/012072.

[18] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," Procedia Comput. Sci., vol. 152, pp. 341–348, 2019, doi: 10.1016/j.procs.2019.05.008.

[19] Kusrini and M. Mashuri, "Sentiment Analysis In Twitter Using Lexicon Based and Polarity Multiplication," 2019, doi: 10.1109/ICAIIT.2019.8834477.

[20] [N. Medagoda, S. Shanmuganathan, and J. L. Whalley, "Sentiment Lexicon Construction," 2015, [Online]. Available: 10.1109/ICNC.2015.7378094.