Journal of Computer Science

Original Research Paper

Mining of Social Media on Covid-19 Big Data Infodemic in Indonesia

Faisal Binsar and Tuga Mauritsius

Information System Management, Binus Graduate Program, Bina Nusantara University, Jakarta, Indonesia

Article history
Received: 22-07-2020
Revised: 14-11-2020
Accepted: 16-11-2020

Corresponding Author:
Faisal Binsar
Information System Management, Binus Graduate Program, Bina Nusantara University, Jakarta, Indonesia
Email: faisal.binsar@binus.ac.id

Abstract: Covid-19 is an unprecedented disaster that is still difficult to contain. During the pandemic, there were a lot of cases that were reported to increase exponentially. In this situation, the dissemination of messages and information was very important. The social media platform has contributed as a channel of communication with unprecedented speed. However, the uncontrolled and irresponsible dissemination of information will result in new problems that can be detrimental to many parties. A lot of information may trigger panic, fear and result in lose hope and even paranoia. The provision of correct and timely information as well as any curative and preventive effort to stop the disease are very important. This study aims to present a method in finding out public opinion through Twitter social media mining in the Indonesian context. We are particularly interested in finding out what people's stance with the pandemic. Some people may fully aware of this threat, but the remaining could be careless about what is going on. It is assumed that this stance could lead to people's obedience to the government's policy on COVID 19 Protocol. It is believed that the opinion is hidden behind the comments in the media. By scrapping the tweets on Twitter during March 2020 using Corona and COVID keywords, we obtained as many as 31,003 tweets. We manually classified the tweets into 3 classes, positive, negative and neutral stances. Predictive models are derived using Support Vector Machine, Random Forest and Naïve Bayes algorithms. Random Forest-based model gives the highest accuracy level as high as 89%, followed by Support Vector Machine as high as 87% and Naïve Bayes as high as 68%. The model can further be used to classify opinions in the future giving valuable information for the government in making policies and steps in overcoming the pandemic.

Keywords: Coronavirus Pandemic, Infodemic Social Media, Sentiment Analysis, Twitter Clustering, Data Science, Machine Learning

Introduction

The World Health Organization (WHO) declared the coronavirus or Covid-19 disease to be a pandemic (WHO, 2020). The pandemic status due to the disease is not yet an antidote to immunity and spread to various regions of the world unexpectedly. The determination of the panic aims to increase global awareness of the spread of a disease, but can also have undesirable effects such as global panic. In six months, coronavirus continues to spread throughout the world, with more than 10 million confirmed cases in 188 countries and more than half a million people have lost their lives (BBC News, 2020). In Indonesia alone at the beginning of July 2020, 91,751 cases were confirmed, 4,459 died, 37,037 people were cured and 50,255 were treated (COVID-19, 2020).

To prevent its spread, the government implements a social distancing policy or often also uses the term physical distancing, which is a non-pharmaceutical policy to prevent the spread of epidemics by keeping a distance between each individual and reducing the frequency of meetings between them. The spread of Covid-19 has frightened the current world situation that is increasingly connected and connected, both physically and digitally.

The Covid-19 pandemic coverage has become the main focus in various media, ranging from newspapers, television, websites and also on social media. The coverage and dissemination of Covid-19 pandemic information on social media have even reached its peak and spread to all levels of society. The widespread of Covid-19 in Indonesian society has raised concerns (Fig. 1). That's because the new type of coronavirus can cause death.

DOI: 10.3844/jcssp.2020.1598.1609

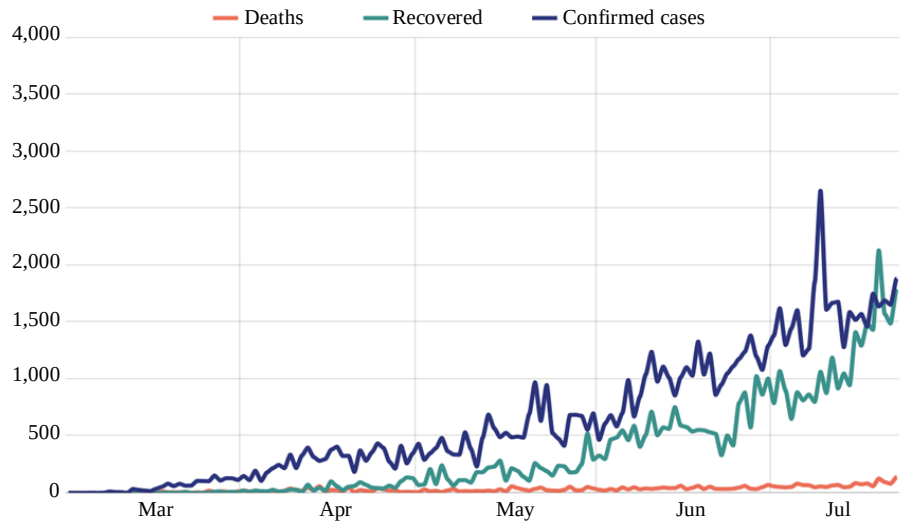


Fig. 1: Cases and handling of Covid-19 in Indonesia (https://www.covid19.go.id, July 22, 2020)

Infodemics on social media occur spontaneously as a result of the Covid-19 pandemic. Everyone can express their feelings, conditions and events experienced or seen from their environment through social media. Everyone can also forward and spread the message obtained. The impact of a public health emergency and this pattern of information dissemination have caused many problems and affected the country and society in the economic, social, religious and other sectors.

Many messages obtained contain important information and are true in fact, many messages are in the form of appeals, socialization from the government. However, messages that unsettle the public and cause fear are no less found through social media. The sender's positive purpose of forwarding the messages received does not always produce positive results for the recipient. Proving the truth of a message or information received is not easy, as well as difficult to realign error messages that are already spread on social media. The reason is that the social media algorithm is directed at attracting people's habits and interests: The emphasis is on liking, not accuracy (MacMillan, 2020).

Social media as an effective way of communication currently contains millions and billions of data messages every day can be seen as a collection of big data. Big data on social media can be used to conduct various analyzes whose results can indirectly be used to reduce the growth rate of Covid-19 infection cases.

Mining data through social media will produce something useful for clear purposes such as getting responses from the community, knowing what the community feels, exploring phenomena that appear on social media during a pandemic (Mukkamala and Beck, 2018) and so on because the output can be used as a reference to warn and isolate people who might have been exposed to the virus. In this technological era, it is only natural that technology and data are used for the common good, especially if in this case the aim is to maintain the health of the general public. Knowledge of positive and negative values of social media opinion is one measure of the feelings that are being felt by the public.

This study uses and compares the ability of the Random Forest algorithm, Support Vector Machine and the Naïve Bayes classifier in classifying public opinion regarding Covid-19 infodemics into neutral, positive, or negative categories. The results of this opinion classification can provide an overview of the psychological state of the community related to the reporting of the Covid-19 pandemic. Negative reporting tends to increase fear in the community. Fear makes a healthy person sick, excessive fear will also reduce physical condition and cause new problems. On the contrary, positive news will still maintain the stability of the condition of the community and avoid things that are not desirable. Providing correct and timely information and supporting healing efforts can prevent stopping the spread of disease.

The presentation of this research begins by explaining the need to understand public opinion related to the coronavirus and assessing opinions into positive, negative and neutral labels in section 2, followed by the use of social mining as an expected solution and discussing other related research. Section 3 explains the proposed methodology and model. Next section 4 explains the analysis of the results obtained from the model used. This study concludes with a note of conclusions and suggestions that can be carried out in subsequent studies in section 5.

DOI: 10.3844/jcssp.2020.1598.1609

Literature Review

Coronavirus Infodemic

Infodemic becomes a word that is quite popular after the outbreak of coronavirus. Information is a keyword (Hua and Shaw, 2020), where stakeholders together with regulations are needed to reduce the impact of false news in this era of information and social media. Although different countries will require different approaches, focusing on the human side and addressing infodemic issues are two important factors for future global mitigation efforts.

During panic storms, digital platforms can encourage the dissemination of problematic information (Jack, 2017). The development of news sites that need to be questioned by publishing stories that are not sourced, cannot be verified or falsified.

Coronavirus was first reported on December 1, 2019, in Wuhan, the case being sporadic throughout December, especially in the latter part of the month. The Wuhan municipal health commission reported that Coronavirus was an unusual disease, which surprised local doctors. Huang *et al.* (2020) reported the unusual case as an epidemic on WeChat social media on December 30, 2019.

With the spread of the Covid-19 epidemic, other major infodemics spread virally throughout the world with repeated drama (Hu *et al*., 2020). Previous evidence shows that the Internet, in essence, can strengthen and convey information quickly throughout the world, causing excessive panic and exacerbating the stigmatization of people at the center of the growing epidemic. The same thing was stated (Hernández-García and Giménez-Júlvez, 2020), that the internet as a large source of health information and can influence its users. However, the information found on the internet often lacks scientific rigor, because anyone can upload content. This factor is a cause for great concern for the scientific community, government and users.

The panic of social media runs faster than the spread of Covid-19 (Depoux *et al*., 2020), misinformation confuses and spreads fear. The impact of media reporting and public sentiment can have a strong influence on the public and private sectors in making decisions to stop certain services including aviation services, not comparable to the actual health needs of the public.

There is a variety of information on social media, including situational related information, information to help and understand the situation during an emergency, information on the number of people affected (Mukkamala and Beck, 2018). This information is useful for the public and authorities to help respond (Martinez-Rojas *et al.*, 2018; Yan and Pedraza-Martinez, 2019). Knowing the type of information and estimating the scale of its distribution can help the authorities to feel the mood of the community, the information gap between the authorities and the public. This knowledge will assist authorities in developing appropriate emergency response strategies (Yan and Pedraza-Martinez, 2019).

Social Media and the Internet

Social media plays an important role in capturing people's thoughts in the representation of their sentences. Twitter, a popular microblog, provides a lot of rich information in it, in terms of short texts posted by users. Such daily and informal sentence processing require special preprocessing techniques to be able to provide understanding and analysis (Ramachandran and Parvathi, 2019).

The use of data from social media platforms has been widely carried out in research. By relying on digital data sources, such as data from cellphones and other digital devices, have special value in outbreaks caused by newly discovered pathogens, official data and reliable forecasts are still scarce. Research (Wu *et al*., 2020) shows the possibility of predicting the spread of the Covid-19 outbreak by combining data from the official flight guide with data on human mobility from the WeChat application and other digital services owned by Chinese technology giant Tencent. Examples of cell phone data use have shown potential in predicting the spatial spread of cholera virus during the cholera epidemic in Haiti in 2010 (Bengtsson *et al*., 2015), while the use of data analysis also shows its effectiveness during the Ebola crisis in the West Africa region in 2014-2016 (Bates, 2017) (Fig. 2).

Many methods of sharing information through giant social media platforms have extraordinary speed, reach and penetration. More than 2.9 billion people use social media regularly. Researchers (Merchant and Lurie, 2020) integrate social media as an important tool in managing a developing pandemic and transforming aspects of preparedness and response for the future. Large volumes of Twitter data streaming (Wang *et al*., 2020) geotag on the flu epidemic provide opportunities for researchers to explore, model and predict trends in flu cases on time.

Big Data and Text Classification

Data mining is the study of collecting, cleaning, processing, analyzing and gaining useful insights from data (Aggarwal, 2015). While the use of big data also plays an important role in pandemic prevention. The utilization of sophisticated computing models using machine learning has shown great potential in tracking sources or predicting the spread of infectious diseases in the future. Therefore it is very important to utilize big data and smart analytics and use them well for public health (Ienca and Vayena, 2020).

Yuan (2020), a staff scientist at the National Institutes, International Division of Epidemiological Health and Population Studies in the US said that big data can help governments effectively estimate the development of a given epidemic and to do so it is necessary to integrate data collection and supervision.

DOI: 10.3844/jcssp.2020.1598.1609

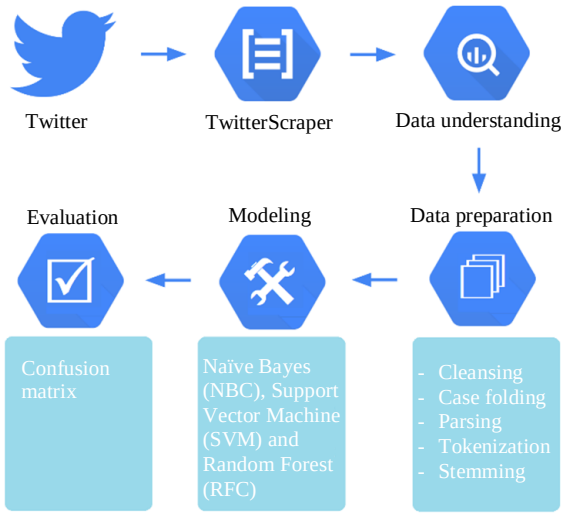


Fig. 2: Stages of research using CRISP-DM

Byrd *et al*. (2016) found the Naïve Bayes classifier to be the most accurate by using location-based Twitter data for sentiment analysis for classification (positive, negative, neutral). The Naive Bayes classification is implemented using Natural Language Processing (NLP) and is trained using an OpenNLP training dataset that includes 100 annotated tweets.

Support Vector Machine (SVM) algorithm has been carried out by many researchers to mine data in predicting some newly discovered viruses. SVM used for surveillance of infectious pandemics has been carried out by (Li and Sun, 2018) to investigate the use of alignment-based and free alignment methods and support vector machines using mononucleotide frequencies and dinucleotide biases to predict host viruses and apply this approach to three data sets: Rabies virus, coronavirus and influenza A.

On social media, SVM has been used to solve many problems, one of which is opinion mining. O'Connor *et al*. (2010) extract tweets to measure people's satisfaction with a product. This approach relies on SVM to divide Twitter posts into positive and negative classes based on the appearance of sentiment words. Likewise, (Zubiaga *et al*., 2011) classifying Twitter posts to summarize trending topics.

Random Forest is a popular machine learning algorithm that is used for several types of classification tasks (Guo *et al*., 2011; Özçift, 2011; Seera and Lim, 2014; Titapiccolo *et al*., 2013). Each tree provides a unit sound, assigning each input to the most likely class label. Can identify non-linear patterns in data, because it can easily handle numerical and categorical data (Titapiccolo *et al*., 2013). Random Forest has also been widely used in classifying public opinion on Twitter (Bahrawi, 2019; Novalita *et al*., 2019; Saleena, 2018; Basha and Somasundaram, 2019).

Methodology

In conducting this research, there are several stages carried out ranging from business processes, preparation of data collection to produce a classification model. The stages used are developing the standard stages of the Cross-Industry Standard Process for Data Mining (CRISP-DM). Mariscal *et al*. (2010) stated CRISP-DM as a standard defacto for the development of data mining and knowledge discovery projects because it is most widely used in data mining development (Mauritsius and Braza, 2019; Husein *et al*., 2020; Oztekin *et al*., 2014).

The CRISP-DM process model provides an overview of the life cycle of a data mining project. CRISP-DM has 6 stages namely business understanding, data understanding, data preparation, modeling, evaluation and deployment.

Data and Data Sources

The tweet data is taken by web scrapping method, which is the process of extracting data from a website, the website, in this case, is the social media website twitter using twitter scrapper.

We use TwitterScraper to replace the weaknesses in the Twitter search API that can only access Tweets that were written in the last 7 days. This is a major obstacle for anyone looking for past data to make a model. TwitterScraper does not provide these restrictions, it can retrieve tweet data at any time in the past. Another disadvantage of the search API on Twitter is that it can only send 180 requests every 15 min. With a maximum number of 100 tweets per Request, this means you can mine 4, 180×100 = 72,000 tweets per hour. TwitterScraper does not limit this amount but is only limited by internet speed or bandwidth.

The data taken is only tweeted in Indonesian with keywords corona and COVID. Data is taken for March 2020 with a limit of 31,003 tweets, or an average of 1000 tweets per day. We chose the month of March because the second week of this month began to enter and the development of COVID in Indonesia, starting this month there was also a lot of news coverage and excitement on social media. Of the many available data fields, we only take one field, namely text with a nomial type, then add one more field, sentiment, which is used as a label with a nominal data type.

Reviews that are pulled from twitter have many attributes including: *Screen_name, username, user_id, tweet_id, tweet_url, timestamp, timestamp_epochs, text, text_html, links, hashtags, has_media, img_urls, video_url, likes, retweets, replies, is_replied, is_reply_to, parent_tweet_id, reply_to_users*. For the assessment in this study, we only use 3 features of the dataset Table 1.

Data Preparation

This process is carried out to produce data that will be used in a machine learning algorithm in a better form than in the original form. The purpose of the preparation process is to eliminate noise, homogenize word shapes and reduce word volume. The stages consist of:

1. Filtering, only take reviews from the public, reviews in the form of news removed from the dataset. The appearance of news or news links in twitter reviews related to Covid-19 is a lot, which is around 30%. The news does not represent a public response, so we do not use it in judging sentiment
2. Labeling, the process of giving neutral, positive, or negative label values to each tweet's data
3. Cleansing, the process of cleaning up reviews of words that are not needed to reduce noise in the classification process. Words that need to be removed such as HTML tags, keywords, emotion icons, hashtags (#), usernames, URLs and emails. The main problem that often arises in language processing in Indonesian on social media is the use of non-standard, messy and abbreviated words that are no longer in accordance with language rules. In the Twitter review, there are a lot of non-standard uses of Indonesian words so that a step is needed to change these words to standard ones. Replacement of standard words is part of cleaning data at the preparation stage. Examples of non-standard words like: 'ga', 'tdk', 'nggak' to 'tidak', 'bgs' to 'bagus', 'gue' to 'saya' and so on. This process is done directly using Python by reading a dataset of non-standard words
4. Case folding is the process of uniforming letters and removing punctuation. In this case, only accept Latin letters between a to z
5. Parsing is the process of breaking a review into a word by analyzing a collection of words by separating the words and determining the syntactic structure of each word
6. Tokenization is the process of separating sentences into meaningful parts and identifying individual entities in the sentence. To find word boundaries, boundaries are identified using spaces and punctuation. The main advantage of Twitter's special tokenization is the complete separation of URLs and hashtags in tweets. When observed in normal techniques, the URL is sorted into many parts and '#' is separated from the words. Hash identification is very helpful in further processing in many applications such as trend detection, opinion mining, event detection and others
7. Stemming is the stage of finding the root word by removing the affixes to a word. The purpose of implementing stemming in this study is to improve performance and reduce the use of system resources by reducing the number of unique words that the system must accommodate. In general, stemming algorithms perform the transformation of a word into a standard morphological representation (known as a stem)

After carrying out these stages, the final dataset is ready to be used as a model in training and testing as many as 19,459 tweets. The data is much reduced because of the sorting of tweet data in the form of news and links from many other websites.

Modeling

Many natural language processing methods can be used to classify social media content. In this study we used three models, namely using the Random Forest Classifier (RFC), Support Vector Machine (SVM) and Naïve Bayes Classifier (NBC). The data used to create the model uses the March tweet review data which is assessed for sentiment manually. Sentiment value consists of 3, namely positive, negative and neutral. The positive sentiment means that the sentence has words that contain more positive meanings, while sentiment indicates that the sentence is composed of more words with negative meaning. While the neutral sentiment is formed because the number of positive meaningful words is equal to the number of negative meaningful words.

The dataset used to make the training model is 70% and the remaining 30% of the data is used to test the model and get an accuracy value from each of the classifying algorithms. The representation of the algorithm model used is explained by the pseudocode shown in Fig. 3.

```
1 : Import library
2 : Load dataset
3 : Text preprocessing
4 : Vectorizer
5 : Split dataset into train and test sets
6 : Run training model : Naïve Bayes, Support Vector & Random Forest
7 : Make predictions : Naïve Bayes, Support Vector & Random Forest
8 : Calculate confusion matrix
9 : Calculate accuracy score
```

Fig. 3: Pseudocode algorithm model

Table 1: Features of the dataset

No	Feature	Explanation
1	*Text*	Contains reviews
2	*Timestamp*	Date and time of review, a review score is calculated per day
3	*Links*	To identify reviews in the form of news links

Evaluation

Evaluation is done using the confusion matrix, namely True Positive rate (TP rate), True Negative rate (TN rate), False Positive rate (FP rate) and False-Negative rate (FN rate) as indicators. TP rate is the percentage of positive classes that have been successfully classified as positive classes, while the TN rate is the percentage of negative classes that have been successfully classified as negative classes. FP rate is a negative class classified as the positive class. FN rate is a positive class that is classified as a negative class.

The values of the Confusion Matrix can be used to calculate the accuracy of the classification of each algorithm. To know the performance of each algorithm, based on the Confusion Matrix this will also be demonstrated through the Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) curves.

ROC shows the performance of classification problems in determining the threshold of a model. With the value of the threshold, it can be seen how many instances are True Positive, False Positive, True Negative and False Negative. which in turn can be known as the best threshold value. From these values, we can calculate the value of True Positive Rate and the value of False Positive Rate. The higher the True Positive Rate and the smaller the False Positive Rate, the better the threshold. AUC is the area under the ROC curve, which illustrates the probability of the sensitivity and specificity variables with a boundary value between 0 and 1. The area under the curve gives an overall picture of the suitability of the model used.

Results and Discussion

Data processing and modeling in this study using Python with the support of several major libraries such as Natural Language Processing (NLP) because the method used is text mining (Byrd *et al*., 2016).

Many Indonesian people's opinions on social media use words that are not standard or informal. In certain situations, the use of non-standard words feels more familiar and already popular. Many also use abbreviations in order to shorten sentences and condense messages in one send. The use of non-standard words and abbreviations is unknown in the process of stemming so that its presence is maintained in full in the sentence. Of course, this is a burden in the training process because it contains more word lists.

Figure 4 shows a word cloud opinion review during March. Wordcloud positive opinion is more dominated by words that have positive connotations and contain expectations such as the appearance of the word "*baik*", "*benar*", "*sehat*", "*selamat*", "*penting*", "*lebih*" and others. While negative wordcloud is dominated by words that add to fear, worry and pessimism such as the appearance of the word "*takut*", "*sakit*", "*mati*", "*salah*", "*kena*", "*mati*", "*kurang*" and others. The picture shows the dominance of many non-standard words in Indonesian, such as the word "*gitu*", "*dah*", "*biar*", "*moga*", "*ga*", "*gak*", "*kayak*" and others. In addition to non-standard words, there are also many abbreviated words such as "*tdk*", "*jg*", "*klo*", "*bgt*", "*dgn*", "*yg*" and others. These words are not known in the steaming process so they still appear in the dataset. An opinion dataset that has discarded a review in the form of news or a link and has been labeled is shown in Fig. 5. The text column is a sentence that has been carried out at the data preparation stage but still found nonstandard words and abbreviations.

DOI: 10.3844/jcssp.2020.1598.1609

(a) Positive (b). Negative

Fig. 4: Wordcloud Indonesian community tweet in March

	text	senti
0	gue sebenernya masuk angin apa corona	neutral
1	indonesia negara tropis kelembaban tinggi jadi...	negative
2	si corona makin jadi jadi laa	neutral
3	pura pura mbahas corona padahal yg nyebarin is...	negative
4	ngga pernah test virus covid ngga lah banyak n...	negative
...	...	...
19454	maret kalimantan timur pasien positif corona ...	neutral
19455	awal bulan nih siap april mop v moga bulan cor...	negative
19456	oke ya habis corona wkwk	neutral
19457	beberapa hari nemu berita ttg tolak masyarakat...	negative
19458	segera pulih dunia segera baik jiwa raga seger...	positive

19459 rows × 2 columns

Fig. 5: Research dataset and manual labeling

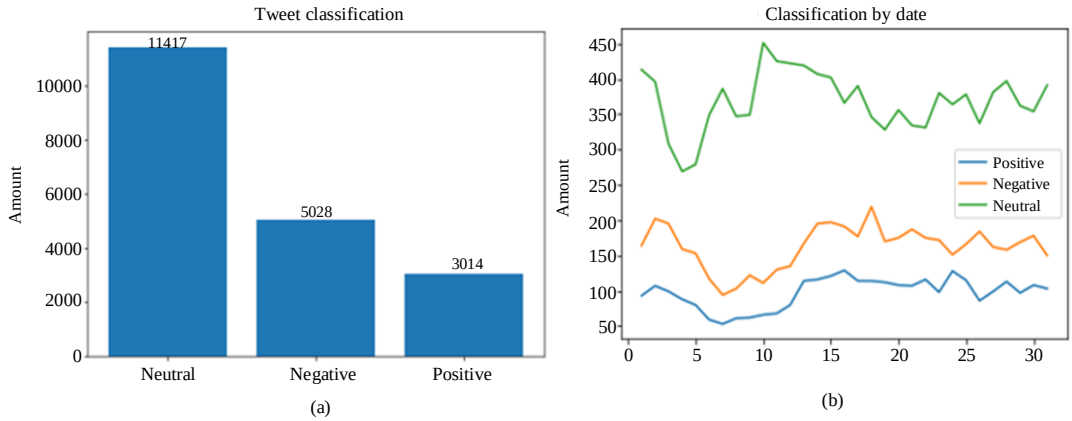


Fig. 6: Classification of manual labeling tweets, (a) the number per label in March, (b) the number of labels per day during March

DOI: 10.3844/jcssp.2020.1598.1609

Sentiment analysis of public opinion in Indonesian is a challenge because the sentence contains many words that are not standard and not under the applicable spelling, as shown in Fig. 5. The use of abbreviations for a word where each person has a different way of abbreviating words so found many abbreviations for the same word. This shows that the data condition is noisy, lost and unstable because it adds very large vocabulary sizes.

Of the 19,459 opinions elected, after labeling each opinion, 11,417 were labeled neutral, 5,028 opinions were negative and 3,014 were labeled positive. The distribution and number of opinion labels per day during March are shown in Fig. 6. The number of negative labeled opinions is more than positive opinions, this shows that fear, worry, or other negative feelings dominate the coronavirus pandemic (MacMillan, 2020) and tend to move be-yond the pandemic (Depoux *et al*, 2020).

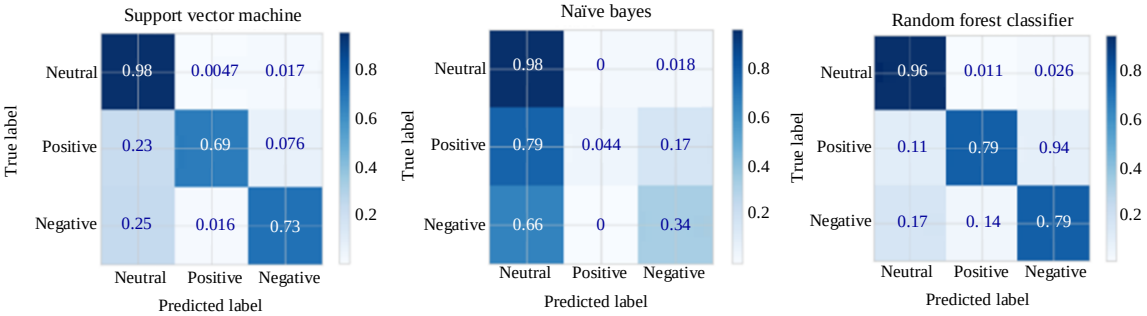


Fig. 7: Confusion matrix of each model

Random Forest Classifier

	precision	recall	f1-score	support
0	0.91	0.96	0.94	1175
1	0.88	0.79	0.84	307
2	0.86	0.79	0.82	464
accuracy			0.89	1946
macro avg	0.88	0.85	0.86	1946
weighted avg	0.89	0.89	0.89	1946

Support Vector Machine

	precision	recall	f1-score	support
0	0.85	0.98	0.91	1705
1	0.94	0.69	0.80	463
2	0.90	0.73	0.81	751
accuracy			0.87	2919
macro avg	0.90	0.80	0.84	2919
weighted avg	0.88	0.87	0.86	2919

Naïve Bayes

	precision	recall	f1-score	support
0	0.67	0.98	0.80	2323
1	1.00	0.04	0.09	608
2	0.69	0.34	0.45	961
accuracy			0.68	3892
macro avg	0.79	0.45	0.45	3892
weighted avg	0.73	0.68	0.60	3892

Fig. 8: Accuracy, precission, recall and f1-score

DOI: 10.3844/jcssp.2020.1598.1609

Three models are used in this study to classify text, all three models produce confusion matrix performance measurement values as shown in Fig. 7.

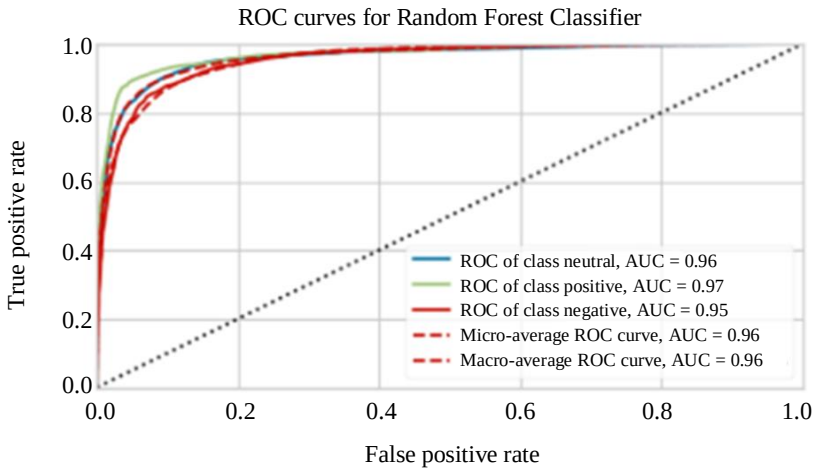
Random Forest Classifier shows the best performance with the best value for the confusion matrix on all labels, followed by the Support Vector Machine model. While the Naïve Bayes Classifier only has a good confusion matrix value for neutral labels.

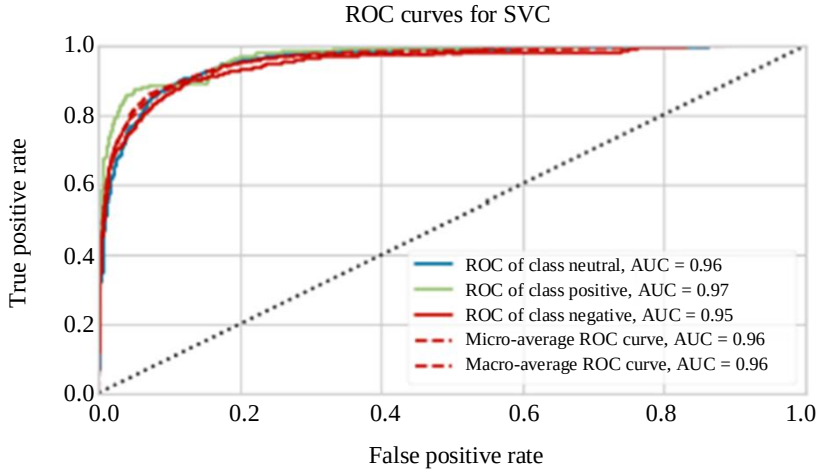
Random Forest Classifier and Support Vector Machine have a very close amount of False Negative and False Positive data (Symmetric) so that the accuracy of both models is very good as a reference for the performance of the model algorithm. On the other hand, Naive Bayes has poor accuracy due to the asymmetrical amount of False Negative and False Positive data.

Based on the confusion matrix that represents predictions and actual conditions (actual) from the data generated by the model used, obtained accuracy, precision, recall and f1-score as shown in Fig. 8.

The accuracy value is obtained from the amount of data that has been successfully classified according to the class of sentiments from the total amount of data classified. The output of the three training models used in this study shows the Random Forest Classifier with the highest accuracy rate of 89%. Then followed by Support Vector Machine with an accuracy of 87%. While the achievement of the accuracy of the Naïve Bayes Classifier is far below the two algorithms, which is 68%. The accuracy of the Random Forest Classifier algorithm in this dataset is better than (Bahrawi, 2019) and (Saleena, 2018), but the Naïve Bayes Classification algorithm in the dataset (Saleena, 2018) has slightly better accuracy.

The performance of each of these algorithms can be known with certainty through the ROC curve and the AUC value shown in Fig. 9. All algorithms show a value close to 1 which means a higher True Positive Rate and a lower False Positive Rate and a good threshold. Of the three models, Random Forest Classifier and Support Vector Machine have the best performance.





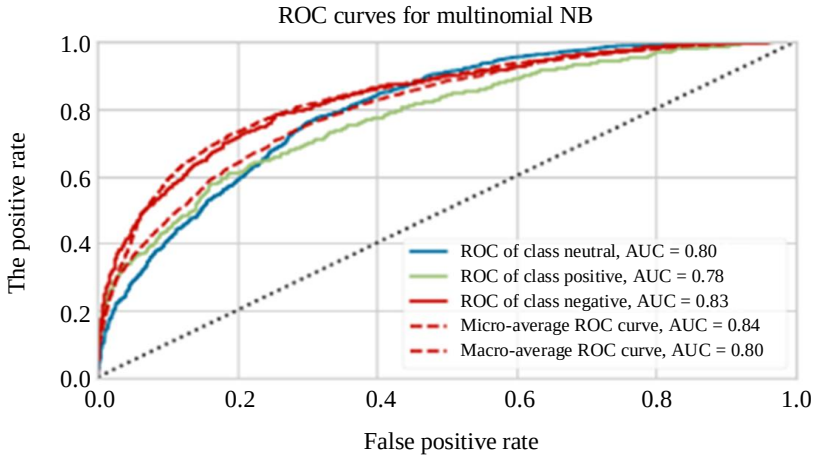


Fig. 9: ROC and AUC of each model

Conclusion

The spread of misinformation about the coronavirus pandemic referred to by WHO as "infodemic," continues. No way can be used to measure many people who believe in false information online. Social media tends to post and disseminate information that further strengthens belief in community groups and often has political objectives. Most people accidentally share information after receiving it through social media and the impact can be fatal. There is also plenty of information that can trigger panic, fear and lead to loss of hope. It is very difficult to convince people on social media that the information they share or post is fake, false, or misleading. Stopping the spread of "infodemic" Covid-19 can be done with care before sharing with others through social media.

Caulfield (2020), director of networked and blended learning at Washington State University-Vancouver, in 2019 suggested using the "SIFT" method (stop, investigate, find, trace) in determining whether the information is seen or read originated from sources that were can be trusted.

The process of classifying social media opinions in Indonesian is very complicated. The general public uses non-standard words. Because of the limited length of text in sending messages, many words are abbreviated, each person has a different way of abbreviating words. The presence of non-standard words and abbreviations will increase the length of the token list with various variations for one word which is the same. Likewise, the existence of negation words does not always contain negative meaning in a complete sentence.

The impact of these characteristics obtained a relatively low level of accuracy in classifying text data as did Random Forest, Support Vector Machine and Naïve Bayes Classifier. Although it produces a fairly good accuracy, the model that was built still made a few mistakes during the process of classifying data that the distribution of sentiments was not balanced. Because using unbalanced data will cause minority class data that is incorrectly classified as majority class data, ultimately making a large difference in value.

We suggest further research specifically opinions in Indonesian to be able to handle various patterns of writing abbreviations and the use of the word negation which can reverse the meaning of the opinions expressed.

Author's Contributions

Faisal Binsar: Defining idea data collection, writing the papers and developing the references.

Tuga Mauritsius: Improving idea and supervision, writing and reviewing.

Ethics

This article is original and contains unpublished material. No ethical issues were involved and the author has no conflict of interest to disclose.

References

Aggarwal, C. C. (2015). Data mining: The textbook. Springer.

Bahrawi, B. (2019). 'Sentiment Analysis Using Random Forest Algorithm Online Social Media Based', Journal Of Information Technology And Its Utilization, 2.2 (2019), 29–33

Basha, C. B., & Somasundaram, K. (2019). A Comparative Study of Twitter Sentiment Analysis Using Machine Learning Algorithms in Big Data. International Journal of Recent Technology and Engineering (IJRTE), 8(1).

Bates, M. (2017). Tracking disease: Digital epidemiology offers new promise in predicting outbreaks. IEEE pulse, 8(1), 18-22.

BBC News. (2020). The Visual and Data Journalism Team, BBC News, 'Coronavirus Pandemic: Tracking the Global Outbreak', Www.Bbc.Com.

Bengtsson, L., Gaudart, J., Lu, X., Moore, S., Wetter, E., Sallah, K., ... & Piarroux, R. (2015). Using mobile phone data to predict the spatial spread of cholera. Scientific reports, 5, 8923.

Byrd, K., Mansurov, A., & Baysal, O. (2016, May). Mining twitter data for influenza detection and surveillance. In Proceedings of the International Workshop on Software Engineering in Healthcare Systems (pp. 43-49).

Caulfield, M. (2020). 'Detect Coronavirus Misinformation in 30 Seconds', WSU Insider, Washington State University, <https://news.wsu.edu/2020/02/27/detect-coronavirus-misinformation-30-seconds/>

COVID-19. (2020). Task Force for the Acceleration of Handling, 'Covid-19 Indonesia'. <https://www.covid19.go.id/>

Depoux, A., Martin, S., Karafillakis, E., Preet, R., Wilder-Smith, A., & Larson, H. (2020). The pandemic of social media panic travels faster than the COVID-19 outbreak.

Guo, L., Chehata, N., Mallet, C., & Boukir, S. (2011). Relevance of airborne lidar and multispectral image data for urban scene classification using Random Forests. ISPRS Journal of Photogrammetry and Remote Sensing, 66(1), 56-66.

Hernández-García, I., & Giménez-Júlvez, T. (2020). Assessment of health information about COVID-19 prevention on the internet: infodemiological study. JMIR public health and surveillance, 6(2), e18717.

Hu, Z., Yang, Z., Li, Q., Zhang, A., & Huang, Y. (2020). Infodemiological study on COVID-19 epidemic and COVID-19 infodemic.

Hua, J., & Shaw, R. (2020). Corona virus (Covid-19)"infodemic" and emerging issues through a data lens: The case of china. International journal of environmental research and public health, 17(7), 2309.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., ... & Cheng, Z. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The lancet, 395(10223), 497-506.

Husein, E., Kaburuan, E. R., & Tuga, M. (2020). 'Mining Data Analysis Using CRISP-DM to Implement Successfully Multipurpose Financing at Astra Credit Companies (ACC) Branch Bogor', International Journal of Advanced Science and Technology, 29.05 (2020), 4497-4506

Ienca, M., & Vayena, E. (2020). On the responsible use of digital data to tackle the COVID-19 pandemic. Nature medicine, 26(4), 463-464.

Jack, C. (2017). Lexicon of lies: Terms for problematic information. Data & Society, 3, 22.

Li, H., & Sun, F. (2018). Comparative studies of alignment, alignment-free and SVM based approaches for predicting the hosts of viruses based on viral sequences. Scientific reports, 8(1), 1-9.

MacMillan. (2020). 'Why the Novel Coronavirus Became a Social Media Nightmare', The Jakarta Post, 2020

Mariscal, G., Marban, O., & Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review, 25(2), 137.

Martinez-Rojas, M., del Carmen Pardo-Ferreira, M., & Rubio-Romero, J. C. (2018). Twitter as a tool for the management and analysis of emergency situations: A systematic literature review. International Journal of Information Management, 43, 196-208.

Mauritsius, T., Braza, A. S., & Fransisca, (2019). 'Bank Marketing Data Mining Using CRISP-DM Approach', International Journal of Advanced Trends in Computer Science and Engineering, 8.5 (2019), 2322–29

Merchant, R. M., & Lurie, N. (2020). Social media and emergency preparedness in response to novel coronavirus. JAMA.

Mukkamala, A., & Beck, R. (2018). The Role of Social Media for Collective Behavior Development in Response to Natural Disasters.

Novalita, N., Herdiani, A., Lukmana, I., & Puspandari, D. (2019, March). Cyberbullying identification on twitter using random forest classifier. In Journal of Physics: Conference Series (Vol. 1192, No. 1, p. 012029). IOP Publishing.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. Tepper School of Business, 559.

Özçift, A. (2011). Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis. Computers in biology and medicine, 41(5), 265-271.

Oztekin, A., Best, K., & Delen, D. (2014, January). Analyzing the Predictability of Exchange Traded Funds Characteristics in the Mutual Fund Market on the Flow of Shares Using a Data Mining Approach. In 2014 47th Hawaii International Conference on System Sciences (pp. 779-788). IEEE.

Ramachandran, D., & Parvathi, R. (2019). Analysis of Twitter Specific Preprocessing Technique for Tweets. Procedia Computer Science, 165, 245-251.

Saleena, N. (2018). An ensemble classification system for twitter sentiment analysis. Procedia Computer Science, 132, 937-946.

Seera, M., & Lim, C. P. (2014). A hybrid intelligent system for medical data classification. Expert Systems with Applications, 41(5), 2239-2249.

Titapiccolo, J. I., Ferrario, M., Cerutti, S., Barbieri, C., Mari, F., Gatti, E., & Signorini, M. G. (2013). Artificial intelligence models to stratify cardiovascular risk in incident hemodialysis patients. Expert systems with applications, 40(11), 4679-4686.

Wang, Y., Xu, K., Kang, Y., Wang, H., Wang, F., & Avram, A. (2020). Regional influenza prediction with sampling twitter data and PDE model. International journal of environmental research and public health, 17(3), 678.

WHO. (2020). 'WHO Director-General's Opening Remarks at the Media Briefing on COVID-19 - 11 March 2020.

Wu, J. T., Leung, K., & Leung, G. M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study. The Lancet, 395(10225), 689-697.

Yan, L., & Pedraza-Martinez, A. J. (2019). Social media for disaster management: Operational value of the social conversation. Production and Operations Management, 28(10), 2514-2532.

Yuan, S. (2020). How China Is Using AI and Big Data to Fight the Coronavirus. Aljazeera.

Zubiaga, A., Spina, D., Fresno, V., & Martínez, R. (2011, October). Classifying trending topics: A typology of conversation triggers on twitter. In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 2461-2464).