IJSA (20xx), Vol. 5, No. 2, 304-313
DOI: https://doi.org/10.29244/ijsa.v5i2p304-313

Indonesian Journal of Statistics and Its Applications
Available at https://ijsa.stats.id
e-ISSN: 2599-0802

Sentiment Analysis on Overseas Tweets on the Impact of COVID-19 in Indonesia*

Tigor Nirman Simanjuntak[1], Setia Pramana[1,2‡]

[1]Directorate of Analysis and Development of Statistics, BPS Statistics Indonesia
[2]Politeknik Statistika STIS, Jakarta Indonesia
[‡]Corresponding author: setia.pramana@stis.ac.id

Abstract

This study aims to conduct analysis to determine the trend of sentiment on tweets about Covid-19 in Indonesia from the Twitter accounts overseas on big data perspective. The data was obtained from Twitter in the period of April 2020, with the word query "Indonesian Corona Virus" from foreign user accounts in English. The process of retrieving data comes from Twitter tweets by crawling the text using Twitter's API (Application Programming Interface) by employing Python programming language. Twitter was chosen because it is very fast and easy to spread through status updates from and among the user accounts. The number of tweets obtained was 8,740 in text format, with a total engagement of 217,316. The data was sorted from the tweets with the largest to smallest engagement, then cleaned from unnecessary fonts and symbols as well as typo words and abbreviations. The sentiment classification was carried out by analytical tools, extracting information with text mining, into positive, negative, and neutral polarity. To sharpen the analysis, the cleaned data was selected only with the largest engagement until those with 100 engagements; then was grouped into 30 sub-topics to be analyzed. The interesting facts are found that most tweets and sub-topics were dominated by the negative sentiment; and some unthinkable sub-topics were talked by many users.

Keywords: covid-19 impact, tweets, sentiment analysis, supervised machine learning, vader.

* Received: Jan 2021; Reviewed: Jan 2021; Published: Jun 2021

1. Introduction

In the midst of the contagion of the Covid-19 outbreak which started in China and then attacked the world globally, various countries are putting all their power and efforts to deal with and stop this pandemic. The efforts in mitigating this pandemic are not only in the ways to cure and prevent the spread of this disease, but also in various aspects such as economic, social, and so on.

In Indonesia, the positive cases have continued to increase since the first case identified in March 2020. This increase was also in line with the increasing ability to test for Covid-19 in Indonesia. Various efforts have been made by the government and society in preventing the transmission of this pandemic, such as physical distancing, social distancing, large-scale social restrictions, healthy and clean lifestyles, flight restrictions, and prohibition of travelling to hometown for Eid Al-Fitr holiday.

Various aspects related to Covid-19 in Indonesia have received a lot of attention and are discussed by many people, both domestic and foreign, on various social channels. Including Twitter, the opinions and talks about Covid-19 in Indonesia are also busy being discussed on this social media microblog. Therefore, this paper tries to capture the opinion sentiment about Covid-19 in Indonesia in various aspects and to conduct analysis to determine the trend of sentiment on tweets about Covid-19 in Indonesia observed from the overseas Twitter accounts on big data perspective.

2. Big Data in Sentiment Analysis

Social Media activities are a digital activity that made up a big data source. The rapid increase in information in social media has demonstrated many opinions, notions, ideas, and comments. One of the most popular microblogs is Twitter, which has managed to attract a large number of users who share opinions, thoughts, and, in general, any kind of information about any topic of their interest. The information that is posted on Twitter frequently contains opinion about products, services, celebrities, events, or anything that is of user's interest (Giachanou & Crestani, 2016).

Due to its increasing popularity, Twitter has recently attracted the interest of many researchers who analyzed Twitter data for a variety of different tasks such as making predictions Bollen et al. (2011), detecting users' sentiment towards different topics Go et al. (2009), detecting users' emotions Mohammad (2012), and detecting irony (Reyes et al., 2013).

Recently, Asri & Mariyah (2019) used Twitter to measure subjective happiness index in Indonesia. In addition, Virati et al. (2019) uses Twitter to investigate the sentiment of unemployment in Indonesia.

Alhajji et al. (2020) performed sentiment analysis on Arabic tweets by using Naïve Bayes machine learning model through the Natural Language Toolkit (NLTK) library in Python. Tweets containing hashtags pertaining to seven public health measures imposed by the government were collected and analyzed. Total tweets analyzed in this study were 53,127. The results show more positive tweets than negative based on the measures applied.

Kaur & Sharma (2020) analyzed the sentiments regarding coronavirus disease (Covid-19). In this research, the API of the Twitter is used for collecting related tweets to the coronavirus, then positive, negative and neutral emotion analyzed by using machine learning approaches and tools powered by NLTK Library and Textblob.

Medford et al. (2020), composed list of hashtags about Covid-19 to search for relevant tweets during a two-week interval from January 14th to 28th, 2020. They were retrieved from the Twitter API and stored as plain text. By using using an unsupervised machine learning, the frequency associated keywords are identified and analyzed such as infection prevention efforts, vaccination, and racial prejudice using an unsupervised machine learning.

Dubey (2020) collected and analyzed tweets from twelve countries. The tweets are collected between 11th march to 31st march 2020, all about the novel Covid-19 disease. The aim of this analysis is to know how individuals in those countries reacting to the outbreaks of the disease. The outcomes of this sutdy shows that most of the people from these societies were thinking positive that the situation would be better, and even though there were also expression of fear and sadness. However, four states especially from the Europe continent demonstrated that they could not trust the situation because of the outbreaks and pandemic over large scale of populations.

Hutto & Gilbert (2014) tested VADER among other highly-regarded sentiment analysis tools such as Linguistic Inquiry Word Count (LIWC), General Inquirer (GI), Affective Norms for English Words (ANEW), SentiWordNet (SWN), SenticNet (SCN), Word-Sense Disambiguation (WSD) using WordNet, and the Hu-Liu04 opinion lexicon, Naive Bayes (NB), Support Vector Machines (SVMs), and Maximum Entropy (ME). The result of the study proved that VADER performed better and remarkable in most cases, producing sentiment polarity.

Having explored various related works and their perspectives, this research attempted to explore the information on most opinions and comments from the overseas Twitter accounts about Covid-19 in Indonesia in a number of particular aspects; and capture sentiment and trend around this topic breaking down into the more specific sub-topics. This research application employed a supervised machine learning with VaderSentiment Package (Valence Aware Dictionary and Sentiment Reasoner).

3. Supervised Machine Learning

Sentiment analysis comprises a multi-step process: a) data retrieval, b) data extraction and selection, c) data pre-processing, d) feature extraction, e) topic detection, and f) data mining process (Hippner & Rentzmann, 2006); Schmunk et al., 2014). In this article, the supervised machine learning is employed by creating a model using annotated data to produce sentiment polarity supported by VaderSentiment Package, a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media (Hutto & Gilbert, 2014).

The Twitter data was collected using the Scrapy package in the Python programming language. This package takes pre-determined information from a website page. In terms of twitter data retrieval, the Scrapy fetches information from the page of search results on the Twitter. This search on Twitter can be carried out based on keywords, hashtags, and account names which can be filtered by the tweeting period. The search results page is then retrieved for tweet information one by one.

The information collected from this twitter retrieval includes the URL of the tweet, the date the tweet was created, the tweet text, username for tweet maker, the number of likes of a tweet, the number of retweets of a tweet, the number of replies to a tweet.

From the data that have been collected, sentiment analysis is carried out for each tweet that has been obtained. Sentiment analysis in this study was carried out using the VaderSentiment package in the Python programming language. The VaderSentiment package help determine the compound value of an English sentence. Compound value itself is a metric that calculates the sum of all normalized lexicon ratings between -1 (most extreme negative) and +1 (most extreme positive) of a sentence. Each English tweet from the collected data will be broken down into a collection of words or also known as tokenization.

After having the words collected, word cleaning process is carried out. This include the sub-processes such as normalization (spelling correction), Stopword Removal (removing words that do not contain content), and Lemmalization (changing words into basic words). Then, after a set of words from a tweet has been cleaned, an assessment of each lexicon is carried out. The result of the total assessment of the set of words is included in the Compound value.

After the compound value of each tweet is obtained, sentiment classification is carried out based on the compound value. The range of the sentiment classification of compound values are as in Table 1.

Table 1: Vader Sentiment Polarity.

Sentiment	Compound Value
Positive	> = 0.05
Neutral	> -0.05 and <0.05
Negative	<= -0.05

4. Sentiment Analysis Discussion

4.1 Overview Result

Of all the tweets crawled from the Twitter with the query “Indonesian Corona Virus” in the period of April 2020, 8,740 tweet texts were obtained, with a total engagement of 217,316. From this number, it can be concluded that the topic of Corona Virus in Indonesia was quite busy being discussed on the Twitter network activities. As we defined in this study, the number of engagements describes the amount of popularity of a tweet through how many times the tweet was retweeted, favourited (liked), and replied. This engagement can indicate how much enthusiasm or interest the topic being discussed (Figure 1).

The negative sentiment amounted to 3,511 (40 per cent of the total 8,740 tweets) on the Covid-19 discussion in April was prominent over the positive tweets (21 per cent or 1,823 tweets), and the neutral sentiment was insignificantly below the negative one, 39 per cent or 3,406 tweets.

To sharpen the analysis, the data was selected only with the largest engagement until those with 100 engagements. Then, the data were categorized into specific sub-topics.

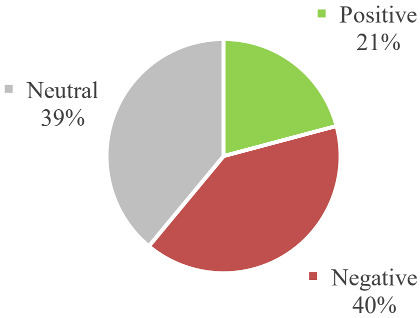

Figure 1: Sentiment Analysis on "Indonesian Corona Virus" Tweets.

4.2 Sub-Topics Classification

To analyze more deeply, the sub-topics that were the most talked about on the Twitter network activities, which are aspects of the Covid-19 topic in Indonesia, a filtering of tweets with the most engagement was carried out. The filtering from total 8,740 tweets resulted in 288 most engaged tweets, from those with the largest engagement (55,321) to the tweets with an engagement of 100 (the least was with 125 engagement). Of these 288 tweets resulted in an engagement of 174,699 which was about 80 percent of the total engagement for all tweets (217,316 engagement).

Then, each of those most engaged tweets was identified and grouped into specific sub-topics which were main interest of the Covid-19 conversation in Indonesia by international user accounts.

The most engaged sub-topic was ‘increase cases’ of Covid-19 in Indonesia, even though it was dominated by the neutral sentiment, this was the most talked and active engagement. The tweets in this group was mostly about number of confirmed cases and death of Covid-19. Issues on how Indonesia handle the Covid-19 cases (sub-topic ‘Indonesia’s handling) ranked at second place with 29,251 engagement that was the sub-topic with the most comments with negative sentiment. Grief at death of medics fighting Covid-19 was the third most sub-topic talked in this observation, amounting to 15,173 engagement and noticeably packed by negative sentiment commonly reflecting sad feeling of loss the medics due to infected by the virus during treating the patients of the Covid-19.

From all the sub-topics (Table 2), the negative sentiment was prominence over the 13 sub-topics. Meanwhile, there were 9 sub-topics with positive sentiment dominance and 8 sub-topics dominated by neutral sentiment.

To see the contribution to the total tweets of most engaged, see Figure 2. The top-ten most engaged sub-topics contributed so significantly to the total tweets at about 82 per cent; even the two most sub topics (Increase in cases and Indonesia’s Handling) were almost half the number of the tweets (of those with most engagement). This proved that these were the most concern of the people overseas in this discussion in April 2020 about Covid-19 condition in Indonesia.

Table 2: Sub-Topic by Number of Engagement and Sentiment Dominance.

Sub-Topic	Engagement	Sentiment Dominance
Increase in cases	55321	Neutral
Indonesia’s handling	29251	Negative
Grief at death of medics fighting Covid-19	15173	Negative
Pessimistic on 2020	14644	Neutral
Unique way of social distancing	10330	Neutral
Testing Ability	8318	Negative
Tabligh Akbar Participant's	8054	Negative
Issues on animal	6669	Negative
Worse than other countries	6264	Negative
Dispute	3029	Negative
Illegal Fishing in Indonesian Waters	2563	Negative
Non-Government contribution	2563	Positive
Flight Cancelled	1626	Positive
Suggestion to Govt.	1504	Negative
Fear of Indonesia	1127	Neutral
Environment	1081	Positive
Issues on Medics	1028	Neutral
Sympathetic Stories	996	Positive
Innovation	960	Positive
Spending behaviour	864	Negative
Better than other countries	792	Neutral
Online wedding reception	464	Neutral
Rise in funeral	436	Negative
Disobey	430	Negative
Appreciation for medics	362	Neutral
Fun facts	271	Positive
International Coordination	181	Positive
Issues on refugees	138	Negative
Deportation of foreigners	135	Positive
Singapore import 4,500 beds	125	Positive

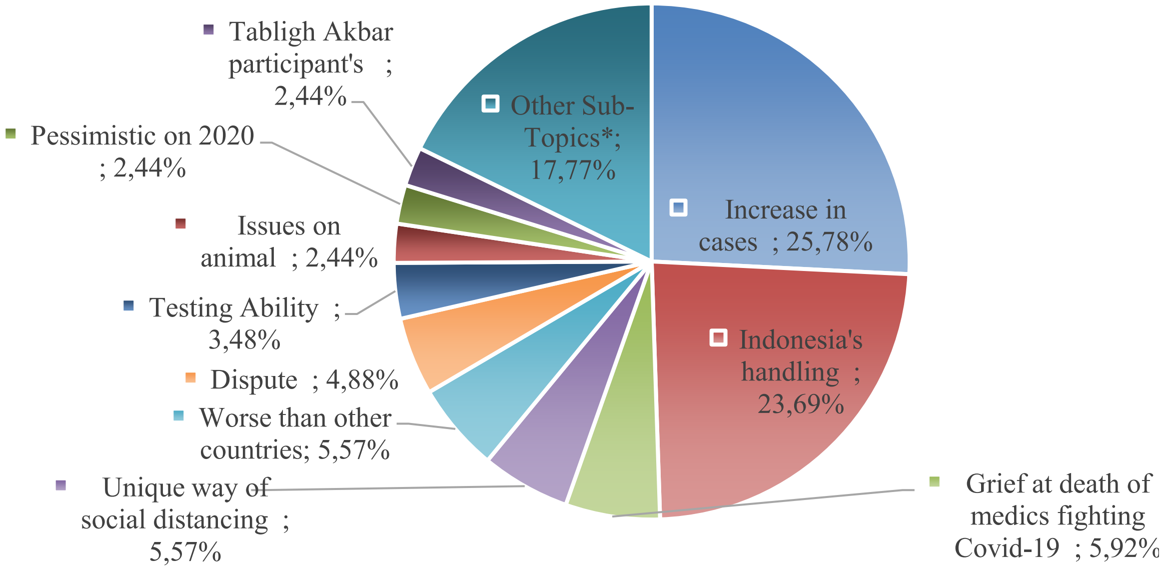

Figure 2: Top 10 Most Sub-Topic by Percentage to Total Tweets of Most Engaged.

The sentiment analysis on the most engaged tweets (the 30 Sub-Topics) is in line with the overview result on the earlier section, where the negative sentiment was the predominant position. In this analysis, the negative sentiment was 45 per cent over the 288 tweets; and followed by the neutral sentiment by 34 per cent and positive by 21 per cent.

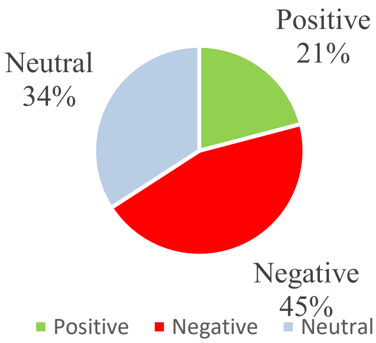

Figure 3: Sentiment Analysis on the Most Engaged Tweets.

4.3 Top Sub-Topic: Increase in Cases of Covid-19

The tweets in this group usually and regularly updated about new cases, new death of the covid-19 or merely regular report of the cases. This also expressed about sharp increase or quick jump of the Indonesian new covid-19 cases. This topic was dominated by neutral sentiment; indeed the positive and negative sentiment were equal in values. This sub-topic was the most commented and retweeted. This demonstrated that most people keep enthusiastic talking about this.

4.4 Sub-Topic Contributing More Negative Sentiment

Indonesia's handling. The tweets in this sub-topic talked about the measures and how Indonesia handle the cases and how it prevent and halt the spread. Most negative sentiment addressed for the slow response of the government over the initial threat of Covid-19 in Indonesia, under-equipped medical personnel fighting the virus, less ability of authorities to restrict the public activity, and for not taking lockdown option.

Grief at death of medics fighting Covid-19. The tweets talked about the loss of doctors, nurse, and other medical/hospital personnel during their struggling to treat the patients. These discussed how sad the story of death of doctors combating the Covid-19 in the frontline that broke many hearts of people. Death of them are perceived to be a great lost as many Twitter accounts sounded.

Worse than other countries. The tweets in this sub-topic talked about the conditions of Indonesia related to the Covid-19 in a number of aspects which are worse than other countries, in terms of death toll rise, highest cases, highest Covid-19 mortality rate, testing rate, and biggest daily spike.

Dispute. The tweets were all about perspective on doubt about data of Covid-19 in Indonesia, due to judgement from particular Twitter users that Indonesia has lack ability to test Covid-19 cases in large number. The arguments for the doubt were seen from the increase of funeral, increasing sale of coffins, testimony from ambulance driver, and data of increasing number of death in all hospital with symptoms like Covid-19 (without Swab result). There were some accounts that consistently speak about the under-reported data of cases and death of Covid-19 in Indonesia.

4.5 Unthinkable Sub-Topics

The interesting findings were a number of unexpected sub-topics - in this perspective - related to the Covid-19 but not directly lead to the main thinking about the virus aspects. They are as follows.

Pessimistic on 2020. This sub-topic discussed about how the year 2020 is difficult, started from the disaster of flood in Jakarta particularly, eruptions of some volcanoes, bushfire, and a number of disasters in other parts of the world.

Issues on animal. Tweets in this group of sub-topic talked about the increase in sale of wild animal due to possibly less intensive check from the authority (probably limitation of operation due to the Covid-19 threat); issues on caged birds, and issues on a group of people burning bats.

Illegal Fishing in Indonesian Waters. This talked about while the world is preoccupied with battling the China-originating coronavirus, China ships were quietly stepped up their incursions into Indonesian waters in suspicious illegal fishing activities.

Environment. This sub-topic said about clearer skies in some Cities in Indonesia; it was assumed to be the effect of limitation of human activities during this pandemic outbreak.

5. Concluding Remark

The tweets about Coronavirus in Indonesia were crowded by the negative sentiment from the foreign users of the Tweeter in the perspective of the Big Data. This was obviously seen in all the tweets crawled with the query "Indonesian Corona Virus" from foreign user accounts in English, where 40 per cent of the total 8,740 tweets were expresses in negative sense, or amounting to 3,511 tweets. The positive sentiment was accounted for 1,823 tweets or around half the negative tweets; and the remained was the neutral sentiment for 3,406 tweets.

This seemed that the aspects of Covid-19 in Indonesia were assessed to be in more negative progress or condition, such as the opinion on how the country is not in quick response to anticipate this outbreak in the earlier time, having high death toll rise, less ability to restrict people activity to halt the spread, less ability in testing rate, under-equipped medical personnel in fighting the virus, and so on.

Breaking down into more specific tweets with the most engagement, the figures were also in the almost similar proportion, where this was also dominated by negative sentiment (45 per cent) and the least was the positive sentiment with the similar figure at 21 per cent.

The information from this method or this research application can be used and further developed to help in search of the new perspective in assessing how broader community have their thought about the condition in many aspects of Covid-19 condition, or any similar subject happening as long as it is talked in the Twitter by the users.

References

Alhajji, M., Al Khalifah, A., Aljubran, M., & Alkhalifah, M. (2020). *Sentiment analysis of tweets in Saudi Arabia regarding governmental preventive measures to contain COVID-19*.

Asri, A. S., & Mariyah, S. (2019). *Subjective Happiness Index based on Twitter in Indonesia*. Retrieved from https://communities.unescap.org/asia-pacific-economic-statistics/apes-2019-featured-papers

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1): 1–8.

Dubey, A. D. (2020). Twitter sentiment analysis during COVID19 outbreak. *Available at SSRN 3572023*.

Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, *49*(2): 1–41.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, *1*(12): 2009.

Hippner, H., & Rentzmann, R. (2006). Text mining. *Informatik-Spektrum*, *29*(4): 287–290.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1).

Kaur, C., & Sharma, A. (2020). *Twitter Sentiment Analysis on Coronavirus using Textblob*. EasyChair.

Medford, R. J., Saleh, S. N., Sumarsono, A., Perl, T. M., & Lehmann, C. U. (2020). An “infodemic”: leveraging high-volume Twitter data to understand early public sentiment for the coronavirus disease 2019 outbreak. *Open Forum Infectious Diseases*, *7*(7), ofaa258. Oxford University Press US.

Mohammad, S. (2012). # Emotional tweets. * *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, 246–255.

Reyes, A., Rosso, P., & Veale, T. (2013). A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation*, *47*(1): 239–268.

Virati, M. Q., Agustiyani, R., Mariyah, S., & Pramana, S. (2019). *Development of a big data analysis system (Case Study: Unemployment statistics)*. Retrieved from https://communities.unescap.org/asia-pacific-economic-statistics/apes-2019-featured-papers