Twitter text mining for sentiment analysis on government's response to forest fires with vader lexicon polarity detection and k-nearest neighbor algorithm

T Mustaqim*, K Umam and M A Muslim

Computer Science Department, Mathematics and Science Faculty, Universitas Negeri Semarang, Indonesia

*Corresponding author: tanzila14@students.unnes.ac.id

Abstract. Opinions about the government's response to forest fires have drawn many opinions from the community. One way for people to express their opinions is to use social media Twitter. This study conducted a sentiment analysis process on the government's response to handling forest fires in Indonesia in 2019 with data sources from Twitter. The analysis was carried out on 6325 datasets written on Twitter on September 20, 2019, and then through the process of pre-processing, automating labeling and classification. The automate labeling process uses a Vader that automatically detects the negative or positive polarity of each data and then goes through the classification process using the KNN algorithm. The test results that were built using rapidminer tools showed an accuracy level of the KNN algorithm of 79.45%, the highest if compared to other classifier algorithms such as decision trees, naïve Bayes and random forests. The sentiment analysis process can almost run automatically without human touch because there is already automated labeling using Vader. Testing sentiment analysis related to the government's response to forest fires can be analyzed using the KNN algorithm and lexicon polarity detection Vader can be done properly.

1. Introduction

Information technology directly plays a role in social life to facilitate and advance all sectors of the daily activities sector, especially to increase the efficiency and effectiveness of the use of resources and accelerate the achievement of objectives [1]. The widespread use of information technology generates a lot of data which if not further processed there is only useless raw data and fills a storage area. Data on the results of human activities contain a lot of valuable information such as habit patterns, opinions and emotional polarity that are useful for certain parties for example to improve services for organizations and decision making related to public opinion [2].

Processing of a lot of data is usually called data mining and in the realm of processing raw data text is called text mining. Sentiment analysis is one of the process of analyzing a collection of corpus texts that aims to analyze the emotional polarity of both negative emotions, neutral emotions and positive emotions [3]. The sentiment analysis process that has been carried out by researchers is using Twitter corpus data. Of all social networks, Twitter has turned into an attractive platform for sharing and communicating with people around the world. People tweet about various issues about news, product brands, politics, etc [4]. Tweet length which only has a limit of around 140 characters makes it easy to share. By writing a tweet, people express their feelings and opinions about different things, such as agreeing or disagreeing, satisfaction or dissatisfaction, their interpretation of various topics or events,
positive or negative feelings [5]. This data source is clearly a great opportunity for business people and academic institutions to conduct research on community thinking.

Sentiment analysis in this research is carried out to find out public sentiments towards Indonesian government opinion on forest fires, whether positive or negative dominant. This study uses the VADER lexicon polarity detection which automatically labels and groups data into neutral, negative or positive categories. Then the training and testing process is carried out using the k-nearest neighbor machine learning algorithm.

2. Related Work

Previous studies relating to sentiment analysis have been carried out, some are used for sentiment analysis of product sentiments and there are also sentiments about the political situation of an area. Research on politics has been carried out by Abdul Malik Zuhdi, Ema Utami and Suwanto Raharjo about Indonesian presidential candidates using the K-NN Algorithm. Their research corpus data were taken crawling from Twitter using the Jokowi and Prabowo keywords in December 2018 with 1000 balanced data between classes. The results of sentiment analysis show that the public has a good opinion of the two candidates, namely positive opinion 300, negative 68, neutral 132 for Jokowi and positive 312, negative 91, neutral 97 for Prabowo then the merging of textual weighting and non-textual weighting can improve accuracy. The accuracy when using textual weighting is 82.50%, when using non-contextual is 60%, while the combination of the two is 83.33% [6].

Sentiment analysis has also been conducted on the reciprocal opinion of the community towards tourism in Oman using the domain specific ontology compiled from Sentistrength, SentiWordNet and Opinion lexicon then combined with semantic sentiment concept analysis. Their research corpus data also came from Twitter with a total of 4432 tweets with criteria using the Oman keyword which was divided into 80% for training data and 20% for testing data. The sentiment analysis results show the highest accuracy level of 85.54% using the concept semantic sentiment analysis [3].

In the next chapter, the steps to collect Twitter corpus data, preprocessing and cleaning Twitter Corpus data, and analyzing Twitter data using VADER sentiment polarity detection and KNN technology will be explained. then we extract sentiment data to find out people's perspectives on public policy in Indonesia.

3. Methodology

3.1. Research procedure

The procedure of this research is explained in Figure 1. The dataset for the analysis was obtained from the 6325 Twitter crawling process with the query "Jokowi" and "Forest Fire".

![Research procedure](image)

Figure 1. Research procedure

3.2. Data collection

We devised a method that can be used to analyze the government's response to forest fires through the Twitter social media dataset. In this research manuscript we collected tweet data from Twitter globally with special queries namely 'jokowi' and 'forest fires'. The two queries are very related because Mr. Jokowi as the president of Indonesia is in control of the handling of widespread forest fires lately. The
method of retrieving data from Twitter in real time is using Twitter streaming API which gives access to data transfer on demand with certain conditions [6].

3.3. Data Pre-Processing

3.3.1. Case folding. Case folding is a method for turning all the letters in a dataset into capital or all small [7]. This is done to facilitate the process of dataset analysis and reduce the amount of memory usage. An example of a folding case is changing the phrase "Forest is in a GREAT FIRE in Kalimantan!" To "for forest is in a great fire in kalimantan!". Case folding helps the lemmatization and stemming process to find a match for each data in the dictionary.

3.3.2. Data cleaning. Every tweet data from social media Twitter usually contains many words and characters that are not useful for the data analysis process [8]. For example there are data tweets such as "RT @sociotalker: what are the similarities of Suharto and Jokowi? Both believe in Wiranto", in that data found useless words or characters such as "RT", "@" and "?". Data cleaning methods combined with regex can detect useless characters and are immediately deleted from the main data to improve the quality of the dataset.

3.3.3. Lemmatization. Non-standard words are often used in communicating and interacting with others. Non-standard words are formed as a result of human interaction itself and sometimes far from the standard rules of the original language dictionary [9]. In the sentiment analysis of nonstandard words is very influential on the results of data analysis calculations.

3.3.4. Automate Data Labeling using VADER. Vader was introduced in 2014 by C.J Hutto and Eric Gilbert whose method of formation was based on a human-centric approach, combining qualitative analysis and empirical validation using human wisdom and judgment [10]. Polarity assessment combines lexical dictionary features with sentiment scores of 5 additional criteria namely exclamation points, uppercase letters, degree of arrangement of words, shifting polarity due to the word "but" and using the tri-gram feature to check the existence of negations.

3.3.5. Remove Stopwords. Very useful such as prepositions, conjunctions, adjectives, slank words, pronouns and much more [11]. These words usually appear together with the main word so that it is not unique and does not have a specific meaning. A list of words that do not contribute too much to analytical text is called a stopword or stoplist.

3.3.6. Stemming. Stemming is the process of removing additional attributes from words such as deleting "mem" and "-kan" from "making" to "making" [12]. In simple stemming is to make the word affect in basic words. Stemming is usually used for information retrieval systems such as search engines and other text analysis.

3.4. Data analysis

3.4.1. Wordcloud. Worcloud is an image composed by many words that implies the contents of the document of analysis [13]. The words are presented in various shapes and colors according to the frequency of occurrence and importance of words in the document. The larger the text form, the more the number of occurrences. Wordcloud is used in sentiment analysis to find out the frequency of dominant words and then draw conclusions according to the topic and condition of the study.

3.4.2. Data Classification with k-Nearest Neighbor (k-NN). KNN is an acronym for k-Nearest Neighbor, which is a classification algorithm whose basic principle is the calculation of the closest distance [14]. KNN algorithm including supervised learning algorithm which requires training data that has been labeled or known in advance. KNN classifies data with unknown labels in accordance with the
training data labels and for the classification process is unlimited because $k$ in KNN can be adjusted to the research needs. KNN algorithm works on vector-shaped data and the conversion process is needed to form vector if the data is not yet in vector form. In the learning phase, this algorithm only stores feature vectors and classifications of learning data. In the classification phase, the same features are calculated for the test data (whose classification is unknown) [15].

4. Results and Discussion
The process of sentiment analysis begins by collecting raw data from Twitter which is taken by crawling using the tweepy program package in the Python programming language. Twitter data retrieval criteria that is data selection in accordance with the query "jokowi" and "kpk". The amount of data obtained was 6325 tweets. Twitter raw data then passes through the case folding process to equalize all forms of characters to facilitate the analysis process [3]. All character shapes are changed to lowercase. From the case folding process Twitter data is then cleared of characters and parts that do not have a contribution in sentiment analysis such as the removal of RT characters, URL addresses and other unique characters.

Twitter data is often found in words that are not in accordance with the rules of standard procedures in the Indonesian language. lemmatization process is needed or the process of changing slang into standard language so that the sentiment results are not ambiguous and more accurate [11]. Twitter data that has been cleared can be labeled using VADER. The automatic labeling process using VADER makes use of the python programming language. Automatic sentiment has been completed then continues to the stopword remover process to facilitate the process of data analysis using wordcloud. Wordcloud is used because it's easier and more interesting in its visual form.

Twitter data is not completely clean for the wordcloud process. The next process is the stemming process. The stemming process is used to change words that have an affix or get out of the basic form to its basic form [13]. This is done so that there is no repeated word analysis and simplifies the wordcloud process. Wordcloud process can be done with the concept of changing words into visual shapes with the most frequency of appearance in different sizes and colors [12]. In this research the wordcloud process is divided into three types namely wordcloud sentiment in neutral, negative and positive. Each is shown in Figure 2.

![Figure 2. Wordcloud Analysis](image-url)
In addition to using WordCloud, this study uses bar charts to show Twitter popular features related to positive or negative sentiment. Each bar graph analysis results are shown in Figure 3.

Figure 3. Hastag Analysis

Table 1 shows a random sample of the results of the classification of Twitter sentiments about government opinion on fires using KNN. Original sentiment is the real sentiment of the results of automatic labeling using vaders.

| No | Tweet                                   | Original Sentiment | Predicted Sentiments (KNN) |
|----|-----------------------------------------|--------------------|----------------------------|
| 1  | banyak pikiran tutup asap pak jokowi bilang | Neutral            | Neutral                    |

KNN algorithm with 6325 tweets data was successfully classified with an accuracy level of 79.45%. the details of the results of the analysis of sentiment analysis are shown in Figure 4.

Figure 4. The results of sentiment analysis

5. Conclusion
In this study an automated labeling process and sentiment analysis were conducted using VADER lexicon polarity detection and K-Nearest Neighbors on raw Twitter data on government opinion on forest fires. Before the data is analyzed, the data is first cleaned through the process of case folding, data cleaning, lemmatization, remove stopword and stemming. Data that is already clean can then be labeled automatically and classified with KNN. The combined work process between VADER lexicon polarity detection and K-Nearest Neighbors can work well on raw Twitter data analysis by producing an accuracy rate of 79.45%.
References

[1] Astuti W T Muslim M A and Sugiharti E, 2019 Sci. J. Informatics 6(1) 95
[2] Safri Y F Arifudin R and Muslim M A, 2018 Sci. J. Informatics 5 (1) 18
[3] Ramanathan V and Meyyappan T, 2019 2019 4th MEC Int. Conf. Big Data Smart City pp 1–5.
[4] Thu P P, 2017 2017 IEEE/ACIS 16th Int. Conf. Comput. Inf. Sci. 2010 209
[5] Gabarron E Dorronzoro E Rivera-Romero O and Wynn R, 2019 J. Diabetes Sci. Technol. 13(3) 439
[6] Tyagi P and Tripathi R C, 2019 SSRN Electron. J. 5 1
[7] Terán L and Mancera J, 2019 Gov. Inf. Q. February 2018 p. 1
[8] Handani S W Intan D Saputra S and Mega R, 2019 J. Phys. Conf. Ser. 1196 1.
[9] Ravi A Khettry A R and Yelandur Sethumadhavachar S 2019 Advances in Computing and Data Sciences, 1, (Springer Singapore) p 403
[10] Hutto C J and Gilbert E E 2014 Proc. 8th Int. Conf. Weblogs Soc. Media, ICWSM 2014
[11] Bahary A F Sibaroni Y and Mubarok M S 2019 Journal of Physics: Conference Series 1192 1
[12] Singh K Shakya H K and Biswas B, 2016 Perspect. Sci. 8 p 570
[13] Wanichavorapong N Che Hussin A R Fadhil Yusof A and Samsuryadi 2019 J. Phys. Conf. Ser. 1196 1
[14] Rezwanul M Ali A and Rahman A, 2017 Int. J. Adv. Comput. Sci. Appl. 8(6) 19
[15] Sharma K and Sambyal A, 2019 Int. J. Inf. Comput. Sci. 6 (2) 1
Catatan:

1. Abstrak maksimal 200 kata, silahkan bisa disesuaikan.
2. Tambahkan referensi pada bagian Results and Discussion
3. Silahkan gunakan Aplikasi untuk pengecekan penggunaan bahasa Inggris. Misalkan grammarly
4. Referensi silahkan diperbaiki sesuai standar IOP (No 1 bisa dijadikan contoh)