ILKOM Jurnal Ilmiah Vol. 15, No.1, 2023, pp.175-185
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

Sentiment Analysis and Classification of Forest Fires in Indonesia

Indra Irawanto [a,1]; Cynthia Widodo [a,2]; Atin Hasanah [a,3]; Prema Adhitya Dharma Kusumah [a,4]; Kusrini [a,5,*]; Kusnawi [a,6]

[a]*Master Of Informatic, University of AMIKOM, Yogyakarta, Indonesia*
[1] *indrairawanto@students.amikom.ac.id;* [2] *cynthiawidodo@students.amikom.ac.id;* [3]*atin.hasanah92s2@students.amikom.ac.id;*
[4] *premaadk@students.amikom.ac.id;* [5] *kusrini@amikom.ac.id;* [6] *khusnawi@amikom.ac.id*
** Corresponding author*

Article history: Received July 27, 2022; Revised August 10, 2022; Accepted March 17, 2023; Available online April 07, 2023

Abstract

Twitter is a well-known social media platform since it allows users to retweet, leave comments, exchange the latest information, and even find out about forest fires. However, no one has processed Twitter data in the form of the topic of forest fires. Despite the fact that this information is incredibly important for determining how much people care about sharing this knowledge and this phenomenon. Hence, one of the efforts in managing Twitter data in the form of text is using NLP (Natural Language Processing) which is now starting to be widely discussed. In addition, the use of word weighting utilizing Vader will also be used in this process. Furthermore, the use classifying process is conducted using 3 kinds of algorithms including Naïve Bayes, Random Forest and SVM (Support Vector Machine). The results of this study, the accuracy obtained from each method has not reached 90%. The Precision, Recall and F1-Score values have also not reached 90%.

Keywords: Sentiment Analysis; Forest fires; Naive Bayes; Random Forest; SVM (Support Vector Machine).

Introduction

Social media is a medium that can disseminate or mislead information compared to traditional or through television broadcasts [1]. One of these social media is Twitter, whose active accounts increase every year. This causes people to connect and communicate with each other in response to news/information [2]. It is possible that Twitter might also discusses forest fire incidents. Moreover, forest fires often occur in Indonesia [3]. This is intriguing because it allows us to gauge how much the American public cares about current events. Moreover, this forest fire is a natural phenomenon that can have a negative impact on nature and anthropogenic ecosystems [4].

Twitter data crawling is the first step to getting sentiment data about forest fires. We can find various languages in the data that we crawl, so the preprocessing process must be carried out as was done in research [5], which retrieved Twitter data on the theme of COVID 2019. Moreover, weighting must be applied to tasks that must be completed prior to classification. This study uses VADER or commonly known as the lexicon. [6] It uses a lexicon that combines lexical dictionary features as a polarity assessment. Sentiment scores of 5 additional criteria, namely exclamation marks, large alphabet, level of word order, polarity shift due to the term "but," and using the tri-gram feature to study negation [7].

Once the text has been labeled, we will classify it using the sentiment analysis. The Nave Bayes technique, Random Forest, and SVM are some reliable classifications that have been demonstrated in numerous research (Support Vector Machine). A popular algorithm that is frequently employed by researchers is Naïve Bayes. The following researchers have used the Naive Bayes method for sentiment analysis research: : [8] analyzing the online store JD.ID, [9] regarding awareness of procedures to prevent COVID 2019. Random Forest is rarely implemented in research on sentiment analysis, although it has recently been investigated to gauge its accuracy. It is used by a number of researchers, including [10], who achieves an accuracy of about 0.829. Moreover, the SVM (Support Vector Machine) approach, whose accuracy is 85%, is also being investigated in sentiment analysis study by [11].

Hence, researchers want to compare the values of the 3 methods namely Naïve Bayes, Random Forest and SVM (Support Vector Machine) to find out the difference in accuracy of the three when using the same data. As for the accuracy will be calculated using the calculation on the confusion matrix. In addition, the researcher also wants to compare the results of classifying sentiment statements which are divided into positive, negative and neutral sentiments.

http://dx.doi.org/10.33096/ilkom.v15i1.1337.175-185

Method

This study adopts descriptive qualitative methods and gathers data from Twitter social media pages. We employ the Python programming language and the provided libraries to carry out this research. The first step in this research project is to conduct a literature review to obtain information or sources about forest fires. Furthermore, this study retrieves tweet data in real-time from Twitter through the Twitter API. Data was collected by taking tweets related to forest fires from Twitter with the keyword "forest fires." Through sentiment, we can find out how users feel about topics or objects [12]. Furthermore, at the data preprocessing stage for data selection and change it to be more structured. At this stage, the cleaning process is carried out to reduce noise and then remove stopwords to remove words that have no meaning [13]. After the data preprocessing process, the next step is translating the data from Indonesian to English.

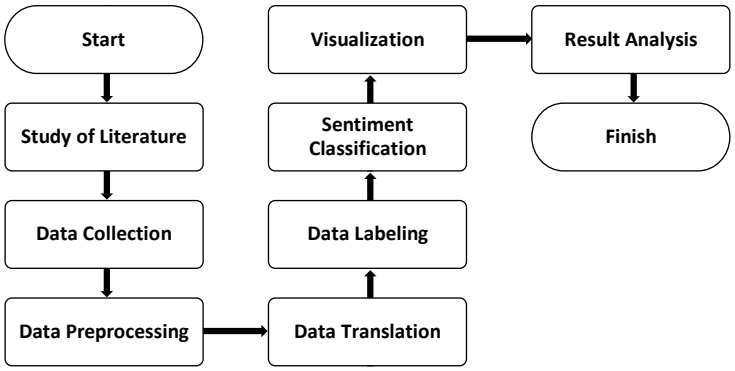

Figure 1. Sentiment Analysis Process Flow

After the translation is complete, the labeling stage can be carried out using lexicon-based comment labeling with VADER to separate positive, negative, and neutral tweets [14]. This lexicon is a collection of words or classes of words that can give weight to the words used in sentences [15]. SVM (Support Vector Machine), Naive Bayes, and Random Forest are three machine learning techniques that will be employed and compared in this study to analyze sentiment at the sentiment classification stage. The model created is assessed using the confusion matrix.

A. Sentiment Analysis

Sentiment analysis or opinion mining is an evaluation model to study public opinion, attitudes and feelings towards any item, product or seller. Opinion Mining mines textual data and evaluates public attitudes towards an object whereas sentiment analysis classifies the sentiments articulated in texts and then examines them [16].

B. Crawling Data Twitter

We use two Twitter API keys, two access tokens (obtained from https://developer.twitter.com/en) and Tweepy Library to crawl and retrieve data from the Twitter using certain trending keywords. We use the term "Forest Fires" as an example for crawling this Twitter data. **Figure 2** illustrates how to use the Tweepy Library, two Twitter API keys, and two access tokens.

```
import tweepy
api_key = "[redacted]"
api_key_secret = "[redacted]"
access_token = "[redacted]"
access_token_secret = "[redacted]"
auth = tweepy.OAuthHandler(api_key, api_key_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth)
topic = 'Kebakaran Hutan'
```

Figure 2. Tweepy Library, API Key and Twitter Access Token

C. Preprocessing Data

Data preprocessing is the most crucial stage because this stage is the process of cleaning and preparing data for analysis. The purpose of preprocessing is to correct the language in conversation, because there are so many non-standard words and local languages being used. The existing data is data from Indonesian to English.

- Case Folding

Changing capital letters to the same letters, specifically lowercase letters, to make them easier to be validate [18]. For instance, "Forest fire engulfed the Park" would change to "forest fire engulfed the park" after going through the case folding stage. The capital letters K, H, and T were converted to their lowercase counterparts.

- Normalization
  Normalization is a stage of the process to clean up the slang language features contained in sentences to be converted into standard language. In addition, this process also cleans URLs, usernames, dates and others [19]. For example, the word "gue" is slang/not standard language, it will be changed to "saya".
- Stopword
  Stopwords remove words that can be ignored in sentences, for example such as adverbs and conjunctions, in this case using the Indonesian stopword dictionary [20]. For example the words "and", "until", "morning", and others.
- Stemming
  Stemming is the process of changing words into basic words from various word formations contained in the sentence [21]. For example "ignore" would be the word "ignore".

D. Google Translate

Language translation is used to translate non-English tweets into English tweets [22]. We benefit from the google translate library to translate sentiments that have gone through the preprocessing stage automatically so they can be processed by VADER/Lexicon in their labeling. The google translate library used, as depicted in **Figure 3**:

```
from googletrans import Translator
translator = Translator()
```

Figure 3. The Google Translate Library

E. VADER atau Lexicon

VADER, often known as Lexicon, is derived from the Greek word lexikon or lexicos [15]. The VADER (Valance Aware Dictionary and Sentiment Reasoner) approach makes it possible to classify texts into negative, positive and neutral sentiment categories [14]. **Figure 4** depicts the Vader Lexicon package in operation.

```
import nltk
nltk.download('vader_lexicon')
from nltk.sentiment.vader import SentimentIntensityAnalyzer
sid = SentimentIntensityAnalyzer()
```

Figure 4. Package Vader Lexicon

F. Naïve Bayes

This is a classification method that relies on Bayes' theorem with a strong (naive) assumption of independence between features. Naive Bayes classification modeling expects that the proximity of certain features (elements) in a class is disconnected from the proximity of several other elements [11]. The Naïve Bayes formula is described in equations (1) and (2) below:

Prior Probability

$$P(H) = \frac{N_j}{N} \quad (1)$$

N_j The amount of data in the class
N The total amount of data

Posterior Probability

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (2)$$

X	Unknown data class	P(H)	Probability hypothesis H (probability prior)
H	certain class X data hypothesis	P(X)	Probability X
P(H|X)	Probability of Hypothesis H based on condition X (posterior probability)	P(X|H)	The probability of X is based on the conditions in the H hypothesis

G. Random Forest

Random Forest is based on the application of bagging to decision trees, with one important extension, besides record sampling [23]. The Random Forest formula is described in equations (3) and (4):

Entropy

$$Entropy(Y) = -\sum_i p(c|Y)\log_2 p(c|Y) \quad (3)$$

Y	Case set
$P(c\|Y)$	The proportion of the value of Y to class c

Information Gain

$$Information\ Gain\ (Y,a) = Entropy(Y) - \sum_v \varepsilon values(a)\frac{Y_v}{Y_a}Entropy(Y_v) \quad (4)$$

$values(a)$	Possible values of the case set a
Y_v	A subclass of Y with class v that is related to class a
Y_a	All values corresponding to a

H. SVM (Support Vector Machine)

SVM (Support Vector Machine) is a supervised learning method used in determining classification. Classification modeling, this method has a more perfect and clearer concept mathematically than the others [24]. SVM can also solve classification and regression problems with linear or non-linear. The formula for SVM (Support Vector Machine), is described in (5), (6), (7), (8), (9), (10) and (11) below:

Matrix Calculations

$$D_{ij} = y_i y_j\left(K\left(\vec{x_i}.\vec{x_j}\right) + \lambda^2\right) \quad (5)$$

D_{ij}	Data matrix elements ij	$K(\vec{x_i}.\vec{x_j})$	kernel function
y_i	The i-th class or data label	λ	Theoretical limit derivatives
y_j	The j-th class or data label		

The n-th data

$$E_i = \sum_{j=1}^{n} a_i D_{ij} \quad (6)$$

$$\delta a_i = min\{max[\gamma(1-E_i), -a_i], c - a_i\} \quad (7)$$

$$a_i = a_i + \delta a_i \quad (8)$$

E_i	The i-th data error value
γ	Learning level
$max_{(i)}D_{ij}$	The maximum value of the hessian matrix diagonal

Finding the bias value (b)

$$b = -\frac{1}{2}[w.x^+ + w.x^-] \quad (9)$$

Decision Calculation

$$h(x) = \begin{cases} +1, if\ w.x + b \geq 0 \\ -1, if\ w.x + b < 0 \end{cases}$$

if the results of the decision calculation ≥0 then the sign h(x) value = +1, belongs to the positive class whereas if the decision calculation results <0 then the sign h(x) value is -1, belongs to the negative class.

$$h(x) = w.x + b \quad (10)$$

atau

$$h(x) = \sum_{i=1}^{m} a_i y_i K(x, x_i) + b \quad (11)$$

I. Accuracy Calculation

The accuracy calculation includes a list of sentiment data, including the dataset's test data, the probability of positive sentiment, negative sentiment [25], and neutral sentiment for each test data, the results of the analysis class with the highest probability, and the accuracy value of the analysis results against the original review sentiment.

In the concept of data mining, accuracy calculations can be obtained using the concept of the Confusion Matrix method. The assessment findings with the Confusion Matrix are in the form of accuracy, precision, recall, and F1-Score. Precision and recall are terms that arise if the right system is designed to be able to show results (retrieve) results in the form of classification, prediction, or search results [10]. Because this study uses 3 labels, it uses Confusion Matrix 3, as shown in **Figure 5**:

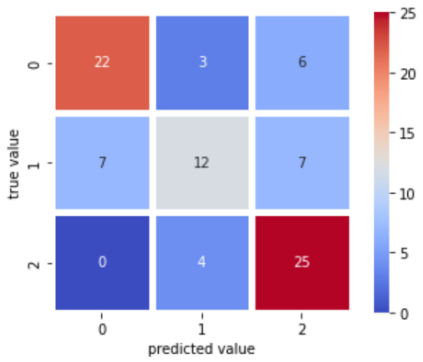

Figure 5. *Confusion Matrix*

The numbers contained in the box in **Figure 5** are the result of the classification carried out by the system, namely the number of original labels that are classified as true or false when compared to the original labels.

Results and Discussion

This section discusses the classification as well as testing of the tweet classification model that was built. After the process of crawling Twitter data obtained as many as 650 tweets. After preprocessing and removing duplicate tweets, a total of 285 tweets were obtained, 98 positive, 102 negative, and 85 neutral. The following steps were taken to analyze sentiment from the tweet data obtained:

A. Data Collection

We use a crawling technique in the data collection process by utilizing the Twitter API Token that has been obtained through https://developer.twitter.com/en. The data we obtained from the Twitter crawl process was 650 tweets. **Table 1** shows the top 3 data and the bottom 3 data obtained:

Table 1. Crawling Results Data from Twitter

Tweet	label
Oak Fire telah memusnahkan 11,900 ekar tanah sejak petang Jumaat dan situasi gagal dibendung. #AWANInews #AWANI745 https://t.co/OrsaorRL3T	1
Gelombang Panas, Ada 3 Titik Kebakaran Hutan di Yunani https://t.co/wbvvTV3gnb #TempoDunia	1
Oak Fire telah memusnahkan 11,900 ekar tanah sejak petang Jumaat dan situasi gagal dibendung. #AWANInews #AWANI745 https://t.co/OrsaorRL3T	1
...	
Kebakaran hutan terjadi di sejumlah negara di Eropa akibat gelombang panas, seperti Prancis, Spanyol, dan Portugal. https://t.co/RlU4k6BCBe #CNNIndonesia https://t.co/VKBdfUy2zf	1

Tweet	label
Kebakaran hutan yang dipicu angin kencang terjadi di kawasan pegunungan Penteli, dekat Ibu Kota Athena pada Rabu pagi, 20 Juli 2022. #TempoDunia https://t.co/UdaS7I5WES	1
Kebakaran hutan terjadi di sejumlah negara di Eropa akibat gelombang panas, seperti Prancis, Spanyol, dan Portugal. https://t.co/RlU4k6BCBe #CNNIndonesia https://t.co/VKBdfUy2zf	1

Table 1 above's data from Twitter crawling results includes terms and sentences that we have highlighted in red. Examples of words that are misspelled or written incorrectly can be found in that section. Moreover, links and hashtags / # are not really necessary for sentiment analysis. So, it is vital to have a preprocessing stage where words are removed, altered, or substituted.

B. Preprocessing

The preprocessing stage, beginning with case folding to change capital letters in tweet sentences into lowercase letters evenly, removing numbers and characters, is shown in Table 2 (the top 3 data). The purpose of this process is to facilitate sentiment validation to the next process.

Table 2. Tweet after Case Folding

Tweet	label
oak fire telah memusnahkan ekar tanah sejak petang jumaat dan situasi gagal dibendung awaninews awani httpstcoorsaorrlt	1
gelombang panas ada titik kebakaran hutan di yunani httpstcowbvvtvgnb tempodunia	1
oak fire telah memusnahkan ekar tanah sejak petang jumaat dan situasi gagal dibendung awaninews awani httpstcoorsaorrlt	1

After Case Folding, then normalize to replace words that don't match EYD to standard words that match EYD. Table 3 shows the words that will be changed later (top 3 and bottom 3) there are as many as 3721 lines in the form of an excel file and this can be added if there are words that are normalized but are not yet in the file:

Table 3. List of Normalized Words

Initial word	Substitute word
singkat	hasil
abis	habis
accent	tekanan
...	
ywdhlh	ya sudahlah
ywis	ya sudah
rp	rupiah

The results of the normalization process are shown in Table 4.

Table 4. Tweets after Normalization

Tweet	label
oak fire memusnahkan ekar tanah petang jumaat situasi gagal dibendung awaninews awani httpstcoorsaorrlt	1
gelombang panas titik kebakaran hutan yunani httpstcowbvvtvgnb tempodunia	1
oak fire memusnahkan ekar tanah petang jumaat situasi gagal dibendung awaninews awani httpstcoorsaorrlt	1

Then do a stopword to remove meaningless words, which are shown in Table 5.

Tabel 5. Tweet after Stopword

Tweet	label
oak fire memusnahkan ekar tanah petang jumaat situasi gagal dibendung awaninews awani httpstcoorsaorrlt	1
gelombang panas titik kebakaran hutan yunani httpstcowbvvtvgnb tempodunia	1
oak fire memusnahkan ekar tanah petang jumaat situasi gagal dibendung awaninews awani httpstcoorsaorrlt	1

The last step is stemming which is done to change sentences into basic words shown in **Table 6.**

Table 6. Tweet after Stemming

Tweet	label
oak fire musnah ekar tanah petang jumaat situasi gagal bendung awaninews awan httpstcoorsaorrlt	1
gelombang panas titik bakar hutan yunani httpstcowbvvtvgnb tempodunia	1
oak fire musnah ekar tanah petang jumaat situasi gagal bendung awaninews awan httpstcoorsaorrlt	1

C. Translate

The next process, the translate stage, utilizes the Google Translate Library feature so that sentiments have already gone through the preprocessing stage. This translated sentiment must be carried out so that it can proceed to the next process, which is the labeling using the VADER technique. The VADER technique is more supportive of English-language texts in the labeling process. The translation results are displayed in **Table 7.**

Table 7. Translate results

Tweet	label
oak fire destroyed acres of land Friday evening dam failure situation cloudinews cloud httpstcoorsaorrlt	1
greek forest fire hotspot httpstcowbvvtvgnb tempodunia	1
oak fire destroyed acres of land Friday evening dam failure situation cloudinews cloud httpstcoorsaorrlt	1

D. VADER atau *Lexicon*

The data labeling stage uses the VADER technique which allows the system to classify information into several sentiment categories, namely negative, positive, and neutral. The determination of the 3 labels is shown in **Figure 6** (positive >= 0.1, negative <= 0 and neutral = 0):

```
df['comp_score'] = df['compound'].apply(lambda c: 'pos' if c >=0.1 else ('neg' if c <0 else 'neu'))
```

Figure 6. Labeling Coding for Compound Values

Below are the results of calculations using VADER and the results of the labeling, which can be seen in **Table 8.**

Table 8. VADER Calculation

Tweet	Negative	Positive	Neutral	Compound	Label
oak fire destroyed acres of land Friday evening dam failure situation cloudinews cloud httpstcoorsaorrlt	0.447	0.553	0	-0.9956	neg
greek forest fire hotspot httpstcowbvvtvgnb tempodunia	0.324	0676	0	-0.34	neg
burn the forests of the united states of america httpstconapueynoy	0	0.763	0.237	0.4215	pos

E. Classification and Accuracy

The classification results were obtained from 285 data divided into 70: 30 (training data: test data), that is, 199 were used as training data, the remaining 86 were used as random test data. Distribution of training data and test data as shown in **Figure 7**:

```
x=data["text"]
y=data["comp_score"]

X_train, X_test, y_train, y_test = train_test_split(x, y,
test_size=0.3, random_state=13,shuffle=True , stratify=y)
```

Figure 7. Library Google Translate

The original label and the classification result label are listed in **Table 9**:

Table 9. Sentiment Analysis Classification Results

Method	Original Label			Classification Result Label		
	Negative	Neutral	Positive	Negative	Neutral	Positive
Random Forest	31	26	29	66	1	19
Naïve Bayes				29	19	38
SVM (Support Vector Machine)				42	25	19

The results of calculations using the confusion matrix are described in **Figure 8**:

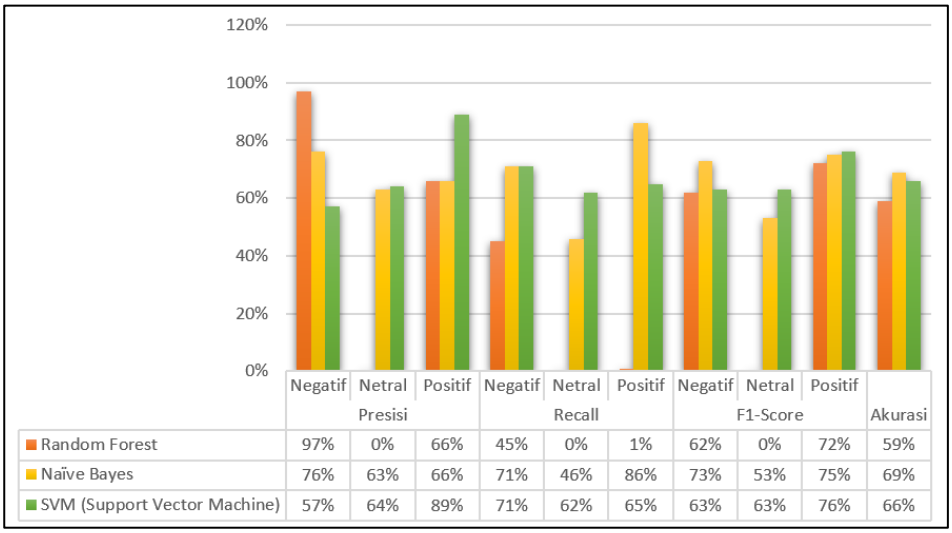

Figure 8. Results of Precision, Recall, F1-Score and Accuracy

F. Application Display

This research is implemented in web form by combining Python and Html. The appearance of the Admin Side Application that has been made is shown in **Figure 9** below:

Figure 9. Display of the Admin Side Application

The Application display on the User side only displays the accuracy results from the crawling – preprocessing data – classification process, shown in **Figure 10** below:

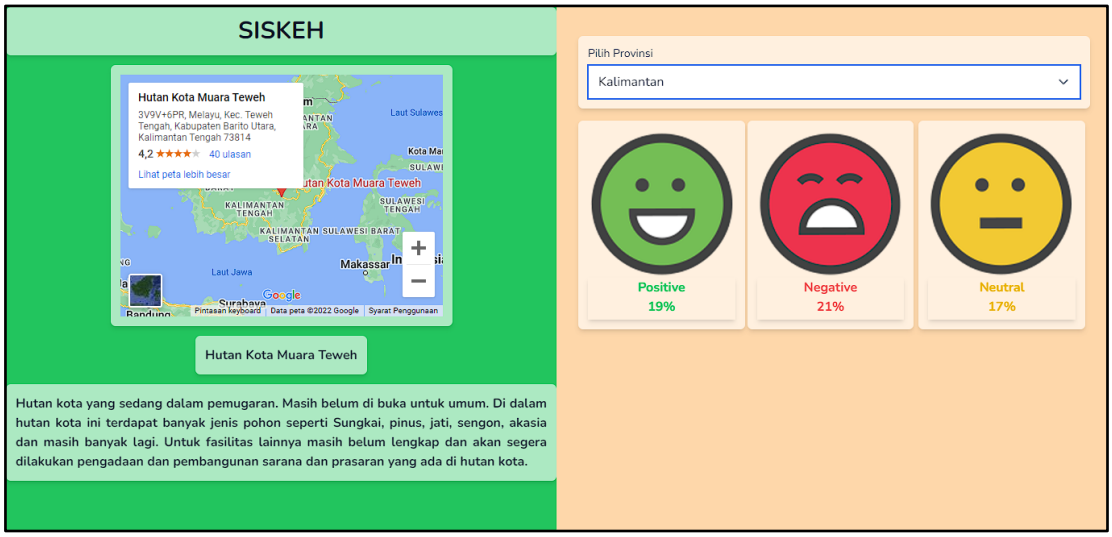

Figure 10. Display of the User Side Application

Conclusion

Based on the results of research that has been done, the accuracy of each method has not reached 90%. The Precision, Recall and F1-Score values have also not reached 90%. In addition, the classification results have been explained in the results and discussion sections. There are several factors that can affect accuracy that does not reach 90%, including: The choice of keywords used throughout the process of crawling data from Twitter may still be less than optimal so that the results obtained are still not optimal. From the keyword "forest fires" only got 650 tweets, after preprocessing, leaving only 285 tweets. In the data preprocessing process, especially the normalization stage, the vocabulary contained in this stage is still incomplete so that there are several slang words that have not been replaced by standard words. Translating data from Indonesian to English, this is a weakness because the translated data must be in the form of original root words. In the labeling process using the VADER lexicon, there may still be errors that can affect the results of the analysis. The distribution of training data and testing data may not be optimal, thus affecting the decision of each method used. This also affects the evaluation of the applied model.

Suggestions for further research are:

1. The choice of keywords needs to be carefully considered and the date range of the tweets you want to crawl can be determined. So that the crawling process can be positioned on several forest fire incidents that have occurred.
2. Retracing the contents of the word list for the normalization stage, so that the registered words can be identified at this stage to be replaced with standard words.
3. Translation from Indonesian to English can use a library other than Google Translate, for example using a deep translator / textblob / goslate or so on. So it can be a comparison of which results are better.

4. The labeling process can use other than Vader Lexicon from NLTK, for example labeling manually but this takes a long time and requires experts in the field or utilizes existing libraries, for example Vader Lexicon from VaderSentiment.
5. The distribution of training data and testing data in our study uses a ratio of 70:30. In future research, you can use a different splitting ratio or use the fold-cross validation method.

References

[1] A. Ristya, C. Chien, and A. Achmad, "Social media sentiment analysis to monitor the performance of vaccination coverage during the early phase of the national COVID-19 vaccine rollout," *Comput. Methods Programs Biomed.*, vol. 221, p. 106838, 2022, doi: 10.1016/j.cmpb.2022.106838.

[2] H. A. Santoso, E. H. Rachmawanto, A. Nugraha, A. A. Nugroho, D. R. I. M. Setiadi, and R. S. Basuki, "Hoax classification and sentiment analysis of Indonesian news using Naive Bayes optimization," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 18, no. 2, pp. 799–806, 2020, doi: 10.12928/TELKOMNIKA.V18I2.14744.

[3] A. Sepriando, H. Hartono, and R. H. Jatmiko, "Deteksi Kebakaran Hutan Dan Lahan Menggunakan Citra Satelit Himawari-8 Di Kalimantan Tengah," *J. Sains Teknol. Modif. Cuaca*, vol. 20, no. 2, pp. 79–89, 2020, doi: 10.29122/jstmc.v20i2.3884.

[4] S. Sakellariou *et al.*, "Remotely sensed data fusion for spatiotemporal geostatistical analysis of forest fire hazard," *Sensors (Switzerland)*, vol. 20, no. 17, pp. 1–20, 2020, doi: 10.3390/s20175014.

[5] M. Hung *et al.*, "Social network analysis of COVID-19 sentiments: Application of artificial intelligence," *J. Med. Internet Res.*, vol. 22, no. 8, pp. 1–13, 2020, doi: 10.2196/22590.

[6] L. Augustyniak, P. Szymánski, T. Kajdanowicz, and W. Tuliglowicz, "Comprehensive study on lexicon-based ensemble classification sentiment analysis," *Entropy*, vol. 18, no. 1, pp. 1–29, 2016, doi: 10.3390/e18010004.

[7] T. Mustaqim, K. Umam, and M. A. Muslim, "Twitter text mining for sentiment analysis on government's response to forest fires with vader lexicon polarity detection and k-nearest neighbor algorithm," *J. Phys. Conf. Ser.*, vol. 1567, no. 3, pp. 8–15, 2020, doi: 10.1088/1742-6596/1567/3/032024.

[8] F. V. Sari and A. Wibowo, "Analisis Sentimen Pelanggan Toko Online Jd.Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi," *J. SIMETRIS*, vol. 10, no. 2, pp. 681–686, 2019.

[9] S. S. Aljameel *et al.*, "A sentiment analysis approach to predict an individual's awareness of the precautionary procedures to prevent covid-19 outbreaks in Saudi Arabia," *Int. J. Environ. Res. Public Health*, vol. 18, no. 1, pp. 1–12, 2021, doi: 10.3390/ijerph18010218.

[10] M. A. Fauzi, "Random forest approach fo sentiment analysis in Indonesian language," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 12, no. 1, pp. 46–50, 2018, doi: 10.11591/ijeecs.v12.i1.pp46-50.

[11] A. Alsaeedi and M. Z. Khan, "A study on sentiment analysis techniques of Twitter data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 2, pp. 361–374, 2019, doi: 10.14569/ijacsa.2019.0100248.

[12] B. Gunawan, H. S. Pratiwi, and E. E. Pratama, "Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes," *J. Edukasi dan Penelit. Inform.*, vol. 4, no. 2, p. 113, 2018, doi: 10.26418/jp.v4i2.27526.

[13] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 1, p. 147, 2021, doi: 10.25126/jtiik.0813944.

[14] E. A. Marwa and A. B. Kristanto, "Analisis Sentimen Pengungkapan Informasi Manajemen : Text Mining Berbasis Metode VADER," *Own. Ris. J. Akunt.*, vol. 6, pp. 2973–2984, 2022.

[15] N. P. Dewi and U. Ubaidi, "Lexical Rule and Lexicon Effect for Part of Speech Tagging Bahasa Madura," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 18, no. 1, pp. 65–72, 2018, doi: 10.30812/matrik.v18i1.332.

[16] P. Balaji, D. Haritha, and O. Nagaraju, "An Overview on Opinion Mining Techniques and Sentiment Analysis," *Int. J. Pure Appl. Math.*, vol. 118, no. 7, pp. 61–69, 2018, [Online]. Available: http://www.ijarcce.com/upload/2014/july/IJARCCE2M s angulakshmi An Analysis on.pdf

[17] E. Chen, K. Lerman, and E. Ferrara, "Tracking social media discourse about the COVID-19 pandemic:

Development of a public coronavirus Twitter data set," *JMIR Public Heal. Surveill.*, vol. 6, no. 2, 2020, doi: 10.2196/19273.

[18] E. H. Muktafin and P. Kusrini, "Sentiments analysis of customer satisfaction in public services using K-nearest neighbors algorithm and natural language processing approach," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 19, no. 1, pp. 146–154, 2021, doi: 10.12928/TELKOMNIKA.V19I1.17417.

[19] N. M. A. J. Astari, Dewa Gede Hendra Divayana, and Gede Indrawan, "Analisis Sentimen Dokumen Twitter Mengenai Dampak Virus Corona Menggunakan Metode Naive Bayes Classifier," *J. Sist. dan Inform.*, vol. 15, no. 1, pp. 27–29, 2020, doi: 10.30864/jsi.v15i1.332.

[20] A. P. Giovani, A. Ardiansyah, T. Haryanti, L. Kurniawati, and W. Gata, "Analisis Sentimen Aplikasi Ruang Guru Di Twitter Menggunakan Algoritma Klasifikasi," *J. Teknoinfo*, vol. 14, no. 2, p. 115, 2020, doi: 10.33365/jti.v14i2.679.

[21] E. Fitri, "Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes, Random Forest Dan Support Vector Machine," *J. Transform.*, vol. 18, no. 1, p. 71, 2020, doi: 10.26623/transformatika.v18i1.2317.

[22] K. Arun and A. Srinagesh, "Multi-lingual Twitter sentiment analysis using machine learning," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 6, pp. 5992–6000, 2020, doi: 10.11591/ijece.v10i6.pp5992-6000.

[23] P. Bruce, A. Bruce, and P. Gedeck, *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python*, 2nd ed. Sebastopol, America: O'Reilly Media, 2020. doi: 10.1080/00401706.2021.1904738.

[24] H. Raza, M. Faizan, A. Hamza, A. Mushtaq, and N. Akhtar, "Scientific text sentiment analysis using machine learning techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 12, pp. 157–165, 2019, doi: 10.14569/ijacsa.2019.0101222.

[25] Y. Nurdiansyah, S. Bukhori, and R. Hidayat, "Sentiment analysis system for movie review in Bahasa Indonesia using naive bayes classifier method," *J. Phys. Conf. Ser.*, vol. 1008, no. 1, 2018, doi: 10.1088/1742-6596/1008/1/012011.