OPEN ACCESS
ISSN 2460-9056
http://socj.telkomuniversity.ac.id/ijoict/

Intl. Journal on ICT
Vol. 8, No. 2, Dec 2022. pp. 66-77
doi: 10.21108/ijoict.v8i2.682

Hoax COVID-19 News Detection Based on Sentiment Analysis in Indonesian Using Support Vector Machine (SVM) Method

Alifia Shafira [1], Yuliant Sibaroni [2], Sri Suryani Prasetyowati [3*]

[123]*School of Computing, Telkom University*
Jl. Telekomunikasi No.1, Bandung, Jawa Barat, Indonesia

[*] alifiashfr@student.telkomuniversity.ac.id

Abstract

The increasing use of technology makes it easier for information media such as news to be disseminated and does not demand possibilities, there is a lot of hoax news spreading. Twitter is one of the media most frequently used by the public to access and disseminate information. This research will focus on detecting Indonesian language COVID-19 news taken from Twitter. Detection of hoax news can be assisted by using sentiment analysis, one of the uses of classification text. Support Vector Machine (SVM) can be used to perform sentiment analysis tasks. After getting the sentiment analysis results, the hoax detection process will use the Bag of Words. Bag of Words is a collection of word dictionaries for weighting words to determine specific labels. The built SVM model succeeded in classifying tweet replies sentiment with an average accuracy of 83.17% with a threshold of 35%. At the same time, the hoax detection process gets the best accuracy of 62.5% with a threshold of -5 or -6.

Keywords: Detection, Hoax, Sentiment Analysis, Support Vector Machine, Twitter.

I. INTRODUCTION

Since the pandemic, almost all activities have been carried out online, which has increased the use of technology and the internet. The existence of this technology makes it easier for people to carry out all activities, one of which is easy to get information. Unfortunately, only some of the information or news is proven true. Many parties misuse this to spread fake news or information, also known as a hoax. Misunderstanding information can trigger many problems.

A news article [1] explained that the Ministry of Communication and Information Technology (Kominfo) said that 9,546 hoaxes had been found spread across various social media platforms on the internet. This data is calculated from August 2018 to early 2022. It is also explained in the same article [1] that the Indonesian people take information sources from social media with a percentage of 73% based on a survey conducted by Kominfo. In the article [2] the Ministry of Communication and Informatics found around 800,000 sites in Indonesia that were indicated as spreaders of fake news. Specific individuals use the Internet to spread negative content that causes anxiety. Based on the data and explanation above, it can be concluded that the problem of spreading hoax news is severe and dangerous. Therefore, research on fake news or hoaxes must be carried out. To prevent issues or wrong perceptions.

Received on 15 Dec 2022. Revised on 26 Dec 2022. Accepted and Published on 31 Dec 2022.

Research on hoax detection has been carried out using various methods. Study [3] used 3,300 Indonesian news, 1,800 non-hoax news, and 1,500 hoax news as training data. An accuracy of 89.27% was obtained using the Support Vector Machine method for 10- folds and an accuracy of 89.48% using the Naïve Bayes method for 5- folds [3]. The expansion feature with Word2Vec can also be used for hoax detection, as was done in research [4]. Study [4] uses the Word2Vec expansion feature with the Logistic Regression algorithm, Support Vector Machine, Random Forest, and a system without the Word2Vec expansion feature method. Using the Word2Vec expansion feature in the Random Forest algorithm increased the accuracy by 1.46% with an accuracy value of 89.53%. Different from the study [4], the study [5] used the TF-IDF weighting feature using the Naïve Bayes Multinomial algorithm to obtain an accuracy of 72.06%.

In research [6] regarding the understanding and analysis of the characteristics of fake news related to sentiment for detecting fake news. That research was verified by comparing several non-sentiment-based fake news detection methods. The classifier method used is LOGIT, SVM-Linear, Decision Tree, Random Forest, XG-Boost, and LSTM_HAN. With the best accuracy results of 86% using the SVM method for sentiment-aware text-only rumor detection. Therefore, this study will use SVM as a model to analyze sentiment toward tweet replies.

In the study [7], hoax detection was based on sentiment analysis with the Naïve Bayes algorithm and lexicon-based analysis. In the study [7], the method used to detect hoaxes is double-checking post comments on Facebook business pages. Where all comments for each post will be analyzed for sentiment, then will proceed to the hoax detection process. The hoax detection process is a double-checking process for post comments based on each post's percentage of comment points.

Based on the previous explanation, in this research, the author will detect hoaxes in tweets crawled from Twitter based on sentiment analysis with the Support Vector Machine algorithm. Different from the study [7], in this research, the hoax detection process will be focused on checking post tweets, not the tweet replies/comments. In this study, there are two different tasks, the first task is sentiment analysis of tweet replies, and the second task is hoax detection of post tweets. The sentiment analysis process for tweet replies is a process for determining which post tweets have the potential to be hoaxes. From these results, the hoax detection process for post tweets uses a Bag of Words for word weighting to determine the "hoax" or "non-hoax" label. This research contributes to detecting hoaxes in the news on Twitter with two different types of data labeling for the process of sentiment analysis and hoax detection using Support Vector Machine.

The paper structure is as follows. Section 2 is an explanation of related research and the methods used. Section 3 is an explanation of the system diagram built for research. Section 4 is an explanation of the results of the research and evaluation. Section 5 is the conclusion of the results of this study.

II. LITERATURE REVIEW

A. Related Works

Research on hoax detection has also been carried out in the study [5]. This study used the Naïve Bayes multinomial method with the TF-IDF weighting feature. By labeling using "hoax" and "non-hoax" labels. Testing in this study is divided into 2 scenarios. The first test scenario uses TF-IDF weighting, which will be tested using N-Grams, and the second test scenario is without using TF-IDF weighting. The highest accuracy result obtained from the first scenario is 72.06% using 80% data train. Whereas in the second scenario, the highest accuracy obtained is 71.65%. With the conclusion that the ratio distribution and the TF-IDF weighting method affect the accuracy value to be obtained.

In 2019 study [7] used the Naïve Bayes and lexicon-based methods to detect hoaxes based on sentiment analysis on Facebook business pages. Where all comments for each post will be analyzed for sentiment, then all posts with a percentage of positive comments lower than 65% will proceed to the hoax detection process.

The hoax detection process calculates points for each comment post based on a dictionary of negative, positive, and hoax/fraud words. From these results, if the point value is lower than 0, the comment is considered negative. Finally, the percentage of total negative comment points for each post will be calculated. The post will be considered a hoax if the percentage is higher than 40%.

In 2020, a study [8] identified hoaxes on Twitter using the Support Vector Machine classification method supported by the Information Gain selection feature and the Term Frequency-Inverse Document Frequency (TF-IDF) weighting feature. The TF-IDF method is a feature extraction for calculating the weight of documents containing frequently appearing terms [8]. This weight value shows how important the term is in a document. Meanwhile, the Information Gain selection feature is used to calculate the effect of a feature on class similarity in a sentence [8]. There were three experiments conducted in this study. First, perform performance tests using SVM by selecting the Information Gain feature and without Information Gain. Second, perform performance tests using SVM with the TF-IDF weighting feature and whether or not to use the Information Gain selection feature with different data presentations. Finally, perform performance tests using SVM with N-grams and whether or not to use the Information Gain selection feature with different data presentations. The three experiments were carried out five times with test data of 10%, 20%, 30%, 40%, and 50%. From this experiment, the highest results were obtained with 10% test data without using the Information Gain selection feature with an accuracy value of 70.21%, while using the Information Gain selection feature, the accuracy value was 95.66% with 0.1 thresholds.

Similar to previous research [8] in research [9] in 2019, researchers also carried out hoax detection on Twitter using the Support Vector Machine and the TF-IDF weighting feature. The difference is that this study used two scenarios. First, changing the composition of the training data and test data. Second, determine what features influence fake news/hoaxes. In the first scenario, the highest accuracy of 78.33% is obtained with a data ratio of 90:10. Whereas from the second scenario, the supporting features are very useful in data labeling and pre-processing plays an important role in the output when the system is tested, lastly data that has been enumerated or labeled greatly influences the results of accuracy.

B. Hoax News Detection

Hoax is information or news that contains things that are uncertain or that are not facts that happened [5]. This hoax is also called fake news. With the ease with which information is spread today, it is also easier for false information to be spread without accountability. To deal with this problem, hoax detection can be carried out in existing news. Usually, in detecting hoaxes, other factors are sometimes used to optimize detection results, such as TF-IDF, sentiment analysis, user credibility, and many other factors.

C. Sentiment Analysis

Sentiment analysis is a process to see whether the content of the text is positive or negative. Usually, sentiment analysis is used to analyze public opinion on a phenomenon. With another explanation, sentiment analysis is a process of extracting emotions or opinions from a text or reading [10]. Before doing sentiment analysis, it is necessary to do data preprocessing. First, change the letters to lowercase because they are capital letters and will have no effect when data processing is carried out. Second, remove special characters from text, such as hashtags, mentions, or URLs. Third, remove multiple whitespaces. Then the last one is tokenization.

D. Support Vector Machine (SVM)

The support vector machine is a classification algorithm that can be described as a dimension that separates two classes of input data. This dimension can be added as a requirement for the data classification process because there need to be more data sets to be separated by just two dimensions. This algorithm aims to find the best boundary to separate the two categories of data, where the closest distance is called the margin, and the

one closest to the margin is called the supporting vector. This algorithm includes the supervised learning algorithm and the most straightforward classification algorithm.

III. RESEARCH METHOD

The hoax detection process that will be carried out in this study is in the form of a sentiment analysis process using the Support Vector Machine (SVM) model to check whether a post tweet has the potential to be a hoax or not. Furthermore, the hoax detection process is through calculating points for each tweet post based on similarities to the Bag of Words. Fig.1 shows a flowchart of the system design for hoax detection in this research.

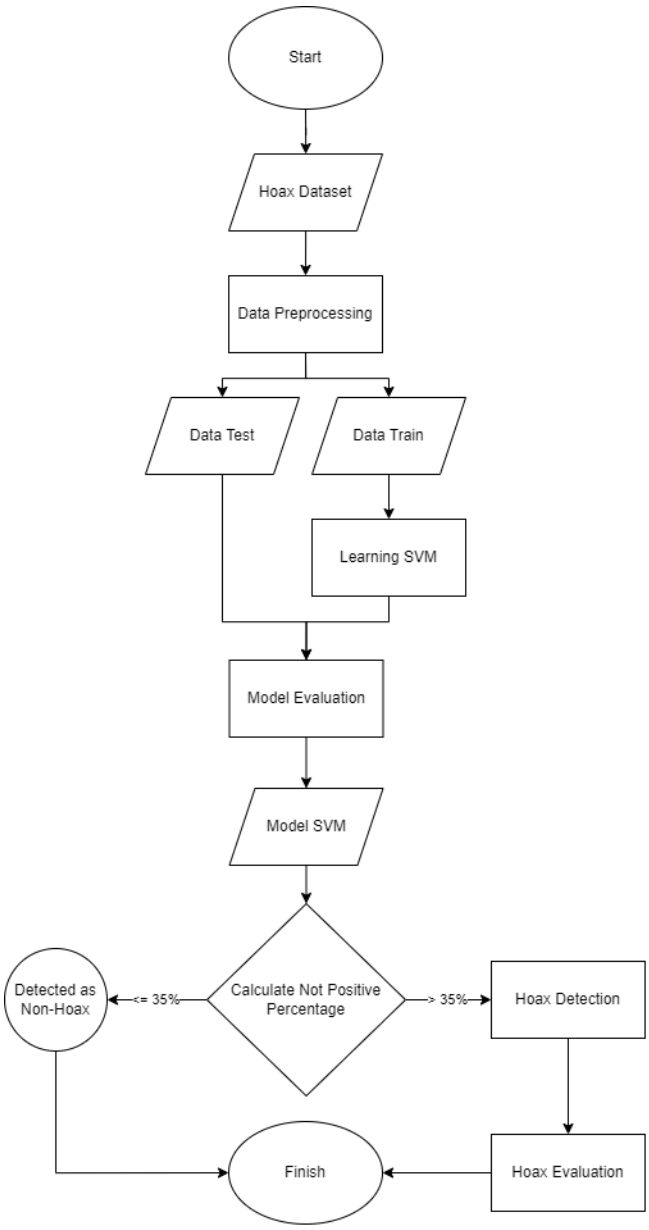

Fig. 1. System Design Hoax Detection

A. Dataset

This research uses a dataset of tweets with the keywords "pandemic covid" and "covid Indonesia". To filter tweets that only contain tweets in Indonesian that discuss COVID-19 because this research will focus on news about COVID-19 that is spreading in Indonesia. This dataset has attributes such as verified, the number of followers, tweet, reply tweet, reply count, retweet, like, count, quote, language, tweet ID, tweet URL, tweet label, and reply tweet label. The example of the dataset format is shown in Table I.

TABLE I
DATASET FORMAT

User	Tweets	Reply Tweets
KawalCOVID19	Indonesia mengumumkan 20.709 kasus baru #COVID19 tgl 9 Aug. Total: 3.686.740 Kasus aktif: 448.508 (-25.725) Sembuh: 3.129.661 (+44.959) Meninggal dunia: 108.571* (+1.475) Org diperiksa: 99.4k (20,8% +ve) - PCR/TCM: 39.2k (36,3% +ve) - Antigen: 60.2k (10,7% +ve) *undercounted https://t.co/SAsPG7GNTV	Ya ampun, sekalian saja tidak usa pake Test, biar 0 kasus. Mau berapapun test nya MR tetap diatas 1000 let's gooo semoga aja ini angkanya beneran, bukan cuma dalam rangka memuluskan agenda penurunan level ppkm Lah gak sampe 100 ribu yg dites, Test turun, yg mati tetap konsisten di angka 1400-1500. Indon lagi
lawancovid19_id	COVID-19 adalah cobaan yang menuntut kita agar kuat untuk bersabar dan berikhtiar melawannya.	Panik gak..panik gak..ya nggak lahh masa panik. Covid19 dan penangannya adalah bentuk kedunguan rezim jokowi yg dibohongi oleh WHO. Ini adalah karma jokowi yg sering bohongi rakyatnya. Waspada variant terbaru virus Corona,jgn sampai terjadi super spreader Lalu kenapa WNA masih saja bisa masuk..a yg bisa menularkan cuma rakyat Kejar target kuartal 2 nih ehem..

B. Data Labeling

Two types of data labels are used: the first is the label for tweets, and the second is the label for the replies to each tweet. For tweets, labeling is in the form of 1 and 0, where 1 means the tweet is a “Hoax” and 0 means it is a “Non-Hoax”. Meanwhile, the tweet replies labels are 1 and 0, where 1 means “Positive” and 0 means “Not Positive”. It needs to be underlined; the labeling used to reply to tweets is "Not Positive", which is different from the label "Negative". Three people carried out the data labeling process. The example of data labeling and the result of data labeling are shown in Table II and III.

TABLE II
EXAMPLE OF DATA LABELING

User	Tweets	Reply Tweets	Label Tweets	Label Reply Tweets
KawalCOVID19	Indonesia mengumumkan 20.709 kasus baru #COVID19 tgl 9 Aug. Total: 3.686.740 Kasus aktif: 448.508 (-25.725) Sembuh: 3.129.661 (+44.959) Meninggal dunia: 108.571* (+1.475) Org diperiksa: 99.4k (20,8% +ve) - PCR/TCM: 39.2k (36,3% +ve) - Antigen: 60.2k (10,7% +ve) *undercounted https://t.co/SAsPG7GNTV	Ya ampun, sekalian saja tidak usa pake Test, biar 0 kasus. Mau berapapun test nya MR tetap diatas 1000 let's gooo semoga aja ini angkanya beneran, bukan cuma dalam rangka memuluskan agenda penurunan level ppkm Lah gak sampe 100 ribu yg dites, Test turun, yg mati tetap konsisten di angka 1400-1500. Indon lagi	0	0 1 0 0
lawancovid19_id	COVID-19 adalah cobaan yang menuntut kita agar kuat untuk bersabar dan berikhtiar melawannya.	Panik gak..panik gak..ya nggak lahh masa panik. Covid19 dan penangannya adalah bentuk kedunguan rezim jokowi yg dibohongi oleh WHO. Ini adalah karma jokowi yg sering bohongi rakyatnya. Waspada variant terbaru virus Corona,jgn sampai terjadi super spreader Lalu kenapa WNA masih saja bisa masuk..a yg bisa menularkan cuma rakyat Kejar target kuartal 2 nih ehem..	1	0 0 0 0

TABLE III
RESULT OF DATA LABELING

Data Type	Hoax	Non-Hoax	Positive	Not Positive	Total
Tweets	34	46	-	-	80
Reply Tweets	-	-	302	1193	1495
Total					1575

C. Data Preprocessing

Data obtained from crawling results are usually unstructured data. Therefore, it is necessary to pre-process the data by processing it so that it is more structured and makes it easier to classify data. The stages in pre-processing data are as follows:

1) Data Cleaning

The data will be cleaned by deleting unnecessary characters, such as characters outside the alphabet a-z, punctuation marks, URLs, hashtags, usernames, and mentions [10].

2) Case Folding

All input data containing capital letters will be converted to lowercase at this stage. This stage is carried out so that all the words in the data input become uniform [9][11].

3) Stopwords Removal

If the sentence contains a stopword, the word will be deleted. Words included in the stopword category are words that are considered not to influence the classification process [11]. This is the example of stopwords "yg", "dg", "rt", "dgn", "ny", “si”, “tdk”. The stopword list uses the NLTK library.

4) Normalization

The purpose of normalization is to change abbreviated words into actual standard words. For example, "udh", "dah", "sdh", and "sudah" all of these words are the same word. This is done so that the model does not recognize it as a different word. The dictionary of normalized words is taken from the source [12].

5) Stemming

The stemming stage is obtaining basic words by removing or eliminating suffixes, prefixes, and inserts at the beginning or end of words [13]. The stemming stage uses the Sastrawi library.

6) Tokenization

A space separates each word in the sentence at this stage. Writing can vary, but the main goal is to cut sentences based on each word that makes up the sentence [10].

D. Sentiment Analysis of Tweet Replies

Sentiment analysis is known as opinion mining to understand the meaning of sentences and phrases [6]. Then proceeding from the data that has been cleaned and pre-processed, sentiment analysis will be carried out. In this study, sentiment analysis uses the Support Vector Machine method. Each reply tweet will be analyzed for sentiment; if the percentage of "not positive" sentiment from the total replies for each tweet is higher than 35%, then it will proceed to the following process, namely hoax detection for the tweet.

Before the classification process, feature extraction is done by weighting the data using the TF-IDF (Term Frequency-Inverse Document Frequency) method. TF-IDF is a method used in weighting word positions in a document [11]. TF states the number of words appearing in the document, while IDF shows how often the word appears in the document [11]. This weighting will make it easier when classifying later.

Support Vector Machine is a supervised learning algorithm that can be used for text classification. SVM is a relatively new technique for making predictions in classification and regression cases [14]. The main idea of the SVM algorithm is to build a hyperplane that can separate areas based on several subsets [15]. The general form of the equation formula is as follows [14]:

$$f(x) = w.x + b \quad (1)$$

$$f(x) = \sum_{i=1}^{m} a_i y_i K(x, x_i) + b \quad (2)$$

x: SVM input data points

w: Parameter hyperlane (the perpendicular line between the hyperlane line and supporting vector)

b: Hyperlane parameter (biased value)

a_i: Data point weight values

K (x,x_i): Kernel function

In equation (2), the supporting vector is meant by a subset of the training set selected as a support vector; in other words, the data x_i corresponds to $a_i \geq 0$ [16].

The data distribution is 80% training data and 20% test data. It is shown empirically that the best results for testing are 20% - 30% of the actual data because the value is close to the v-fold cross-validation, which is considered to give the best results [17].

From the classification results obtained, it is necessary to evaluate using a confusion matrix. Confusion matrix is a tool for conducting analysis usually used in supervised learning to see test results from predictive models [8]. The following is a confusion matrix table shown in Table IV and its mathematical formula:

TABLE IV
CONFUSION MATRIX

		Actual Values	
		Positive	**Negative**
Predict Values	**Positive**	TP	FN
	Negative	FP	TN

To calculate the values of the four classification accuracy parameters, a confusion matrix table is needed, which contains True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) values [17]. This value will be used to get accuracy, precision, and recall [17].

$$Precision = \frac{TP}{TP + FP} \times 100 \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \times 100 \quad (4)$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \times 100 \quad (5)$$

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall} \times 100 \quad (6)$$

Precision shows the level of accuracy between the results of user classification and the results of system classification, recall shows the level of success of the system in finding information, accuracy shows the success rate of the system doing the classification, and F-Measure is the average result of precision and recall [18].

E. Hoax Tweet Detection

The technique used in this study is a text classification system using a machine learning-based approach [19]. With data that is crawled and then pre-processed and then classified until the results of sentiment analysis are obtained. The results, in the form of a percentage of positive or non-positive values, can be analyzed; if the percentage of non-positive values is higher than the positive value, then the data will be checked for the hoax detection process. Meanwhile, if the percentage of positive values is higher than the non-positive value, the data is not a hoax and will not proceed to the hoax detection stage.

The data used in the hoax detection stage is tweets with a percentage of non-positive sentiment higher than 35%. At this stage, a word dictionary will be needed containing words including fraud/hoax, positive, or negative as a reference. An example of a word dictionary is shown in Table V.

TABLE V
WORD DICTIONARY (BAG OF WORDS)

Hoax	Positive	Negative
Curang, penipuan, kedok, illegal, menyimpang, duplikat, sesat, palsu, bohong.	Agung, aman, andal, komitmen, martabat, pro.	Anarkis, cemooh, cacat, diskredit, dilecehkan, iri, keji, munafik.

Each word that has similarities with the word dictionary will get one point. Based on this, equations can be made, such as [7]:

$$C_S = P_S - F_S - N_S \quad (7)$$

With an explanation, C_S is a point for tweet news, P_S is a point for positive words, F_S is a point for a hoax, and N_S is a point for negative words [7].

From the results of the equation above, if the tweet news point has a value higher than 0, then the tweet news is positive and will be considered not a hoax. Meanwhile, if the tweet news point is less than 0, the tweet news is negative, which is considered a hoax.

IV. RESULTS AND DISCUSSION

To evaluate the results of sentiment classification, the accuracy value is used as a reference for the system's success rate in carrying out the classification. Several accuracy values will be obtained from each tweet.

Accuracy comes from the total percentage of positive tweet replies from each tweet, where each tweet has many replies. The model used in the system is a Support Vector Machine (SVM) with parameters of the linear kernel and C as a regularization with a value of 1. The results obtained an average accuracy of 83.17% from the total data test. With the results obtained, the model succeeded in classifying sentiment. Where the threshold limit used is 35%. Based on the threshold determination test, the 35% threshold gives the best accuracy results. Table VI shows the trial for determining the best threshold.

TABLE VI
TOTAL POSITIVE AND NEGATIVE PERCENTAGE BASED ON SENTIMENT PERCENTAGE

% Sentiment	% Total Positive	% Total Not Positive	Total Positive	Total Not Positive
>= 10%	36,25%	93,75%	29	75
>= 20%	27,5%	92,5%	22	74
>= 30%	18,75%	90%	15	72
>= 40%	17,5%	90%	14	72
>= 50%	13,75%	86,25%	11	69
>= 60%	11,25%	82,5%	9	66
>= 70%	10%	81,25%	8	65
>= 80%	7,5%	73,75%	6	59
>= 90%	6,25%	63,75%	5	51
100%	3,75%	46,25%	3	37

As a comparison, in research [8], modeling made using SVM with TF-IDF feature extraction obtained an accuracy of 70.21%. Using a Twitter dataset based on hashtags that are trending in a specific time period. From these results, this study succeeded in making the SVM model with higher accuracy results, i.e. 83.17%. However, in the research [8] the scenario using information gain, there is an increase in accuracy, i.e. 95.66%. There are several reasons why the research accuracy was higher in this scenario. The main reason is that this scenario uses information gain feature selection which helps increase the accuracy value. Then the size of the dataset used is larger. It is also mentioned in the study [8] used feature acquisition, which is used in the system performance process that gives better accuracy.

Sentiment classification will provide an index of which tweets will proceed to the hoax detection stage. The accuracy of the total points calculated for each tweet based on checking with a dictionary of positive, negative, and hoax words with threshold testing 0 is 43.75%. An experiment was carried out to improve accuracy to check what threshold gave the best accuracy. The following are the accuracy results obtained based on the changes in the threshold.

TABLE VII
ACCURACY RESULTS BASED ON THRESHOLDS

Thresholds	Accuracy
0	43,75%
-1	50%
-2	50%
-3	50%
-4	50%
-5	62,5%
-6	62,5%
-7	56,25%
-8	56,25%
-9	56,25%
-10	56,25%

The experimental results found that the -5 and -6 thresholds provide the best hoax detection accuracy, with an accuracy of 62.5%. Then every tweet point that has been calculated will be considered a hoax if the total point less than -5 and considered not a hoax if the total point is higher than -5.

In addition to the experiments above, to improve accuracy, word dictionaries were also added. The word hoax is added based on the words that appear in the tweet data labeled a "Hoax". Before that, each tweet is pre-processed, and the frequency value of the word that appears is calculated. After that, which words will be added to the dictionary will be checked manually. From the results of the accuracy of the hoax detection process that has been obtained, performance still needs to be improved so that the accuracy results can increase.

V. CONCLUSION

This study aims to identify hoaxes based on sentiment analysis using the Support Vector Machine method. Models are built using libraries provided by Scikit Learn. The dataset used comes from Twitter, containing 80 tweet post and 1495 tweet replies.

Before the hoax detection process, each reply tweet will be classified as a sentiment. The built SVM model succeeded in classifying reply tweets with an average accuracy of 83.17%. Furthermore, every tweet with a non-positive percentage higher than 35% will proceed to the hoax detection process. Hoax detection is obtained by calculating points for each tweet based on similarities with the word dictionary. This hoax detection process gets the best accuracy of 62.5% with a threshold of -5 or -6.

This research still needs improvement in terms of accuracy. Because there are factors that affect research, such as the number of datasets that can be enlarged and there are still imbalanced data. The dictionary for the word hoax can also be enlarged so that the model's reference to the hoax word to be studied is better.

REFERENCES

[1] "Hingga Awal 2022, Kominfo Temukan 9.546 Hoaks di Internet - Bisnis Tempo.co." https://bisnis.tempo.co/read/1558213/hingga-awal-2022-kominfo-temukan-9-546-hoaks-di-internet (accessed Dec. 28, 2022).

[2] "Kementerian Komunikasi dan Informatika." https://www.kominfo.go.id/content/detail/12008/ada-800000-situs-penyebar-hoax-di-indonesia/0/sorotan_media (accessed Dec. 28, 2022).

[3] I. Fredy Ferdiansyah, "LAMPIRAN 62 Hoax Detection Analysis Using Support Vector Machine, Naive Bayes, Random Forest and K-Nearest Neighbor Algorithm On Covid-19 Vaccine News On Twitter."

[4] F. Ismayanti and E. B. Setiawan, "Deteksi Konten Hoax Berbahasa Indonesia di Twitter Menggunakan Fitur Ekspansi dengan Word2Vec."

[5] C. S. Sriyano and E. B. Setiawan, "Pendeteksian Berita Hoax Menggunakan Naive Bayes Multinomial Pada Twitter dengan Fitur Pembobotan TF-IDF."

[6] O. Ajao, D. Bhowmik, and S. Zargari, "Sentiment Aware Fake News Detection On Online Social Networks," 2019.

[7] S. Nasrin, P. Ghosh, S. M. Mazharul, H. Chowdhury, and S. A. Hossain, "Fraud detection of Facebook business page based on sentiment analysis," 2019. [Online]. Available: https://www.researchgate.net/publication/331036609

[8] I. Mumu, M. #1, and E. B. Setiawan, "The Effect of Information Gain Feature Selection for Hoax Identification in Twitter Using Classification Method Support Vector Machine", doi: 10.21108/indojc.2020.5.2.499.

[9] A. Fauzi, E. B. Setiawan, and Z. K. A. Baizal, "Hoax News Detection on Twitter using Term Frequency Inverse Document Frequency and Support Vector Machine Method," in *Journal of Physics: Conference Series*, May 2019, vol. 1192, no. 1. doi: 10.1088/1742-6596/1192/1/012025.

[10] A. M. Pravina, I. Cholissodin, and P. P. Adikara, "Analisis Sentimen Tentang Opini Maskapai Penerbangan pada Dokumen Twitter Menggunakan Algoritme Support Vector Machine (SVM)," 2019. [Online]. Available: http://j-ptiik.ub.ac.id

[11] S. K. Dirjen *et al.*, "Terakreditasi SINTA Peringkat 2 Hoax Detection on Twitter using Feed-forward and Back-propagation Neural Networks Method," *masa berlaku mulai*, vol. 1, no. 3, pp. 648–654, 2017.

[12] "Sentimen-Analisis-Bahasa-Indonesia-Menggunakan-Metode-Support-Vector-Machine/normalisasi.csv at main · Inazuna/Sentimen-Analisis-Bahasa-Indonesia-Menggunakan-Metode-Support-Vector-Machine." https://github.com/Inazuna/Sentimen-Analisis-Bahasa-Indonesia-Menggunakan-Metode-Support-Vector-Machine/blob/main/normalisasi.csv (accessed Dec. 15, 2022).

[13] T. Trisna Astono Putri, H. S. Warra, I. Yanti Sitepu, and M. Sihombing, "Analysis And Detection Of Hoax Contents In Indonesian News Based On Machine Learning," 2019.

[14] R. Mahendrajaya, G. A. Buntoro, and M. B. Setyawan, Analisis Sentimen Pengguna Gopay Menggunakan Metode Lexicon Based Dan Support Vector Machine," 2019. [Online]. Available: http://studentjournal.umpo.ac.id/index.php/komputek

[15] A. B. Prasetijo, R. R. Isnanto, D. Eridani, Y. A. A. Soetrisno, M. Arfan, and A. Sofwan, "Hoax detection system on Indonesian news sites based on text classification using SVM and SGD," in *Proceedings - 2017 4th International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE 2017*, Jul. 2017, vol. 2018-January, pp. 45–49. doi: 10.1109/ICITACEE.2017.8257673.

[16] A. S. Nugroho, A. B. Witarto, and D. Handoko, "Support Vector Machine-Teori dan Aplikasinya dalam Bioinformatika 1," 2003. [Online]. Available: http://asnugroho.net

[17] R. Taqiuddin, F. A. Bachtiar, and W. Purnomo, "Opinion Spam Classification on Steam Review using Support Vector Machine with Lexicon-Based Features," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, Nov. 2021, doi: 10.22219/kinetik.v6i4.1323.

[18] B. E. Pasaribu, A. Herdiani, and W. Astuti, "Deteksi Fake Reviews Menggunakan Support Vector Machine."

[19] U. Munirul, M. Mahendra Alvanof, and R. Triandi, "Analisa Dan Deteksi Konten Hoax Pada Media Berita Indonesia Menggunakan Machine Learning."