Syntax Literate: Jurnal Ilmiah Indonesia p–ISSN: 2541-0849
e-ISSN: 2548-1398
Vol. 7, No. 6, Juni 2022

DETECTION OF NEGATIVE CONTENT (HOAX) ON MICROBLOG DATA THAT CONTAINS COVID-19 INFORMATION

Putra Tresna Linge, Alfan Farizki Wicaksono
Magister Teknologi Informasi Universitas Indonesia Jakarta, Indonesia
Email: putra.tresna@ui.ac.id, alfan@cs.ui.ac.id

Abstract

Over the past few years, the amount of information dissemination has increased, especially since the advent of social media. Among the information circulating, there is information that includes negative content or hoax that have a bad impact such as the emergence of divisions due to incorrect information. Based on the 2018 Kominfo performance report, Twitter social media is the largest contributor to the spread of hoax. To reduce the impact of the spread of hoax, a method is needed to detect hoaxes on Twitter so that prevention can be done such as taking down tweets that are hoax. The purpose of this research is to develop a model that can detect negative content (hoax) automatically and also see the correlation between hoax content and sentiment orientation. The results of this study are a machine learning-based model using a decision tree algorithm with an accuracy of 97.2% with a precision value of 85.4, recall of 81.4, and f1-score 93 and the model. In addition, the results of the analysis show that tweets that are hoax as a result of model identification are dominated by positive sentiment orientation, which is 52.64% of the total data identified as hoax

Keywords: Hoax Detection, Twitter, Sentiment Orientation Classification, Machine Learning, Teks Analysis

Introduction

Currently, the exchange of information takes place in a short time and massive amounts. Negative content (hoax) can have various bad effects on the information circulating. Examples of bad impacts are the split during the general election era, where each faction creates hoaxes to bring down the other faction(Juditha, 2019),(Sutantohadi & Rokhimatul Wakhidah, 2017). In addition, hoaxes can also cause terror or fear, as happened some time ago related to information on COVID-19 which led to "panic buying" behavior (Alamsyah, 2020; Somantri, 2020; Wardani, 2017). Covid-19 is a new variant of the virus where information in the form of facts is still minimally known by many people. This has caused several hoaxes, especially those related to COVID-19. Hoaxes about covid-19 that most often appear are related to the presence or absence of covid-19 and the covid-19 vaccine.

Reports for handling hoaxes in recent years have found that there has been an increase in the spread of hoaxes in the health sector on social media, especially on Twitter. To overcome this, the government made efforts to deal with hoaxes by

How to cite: Putra Tresna Linge, Alfan Farizki Wicaksono (2022) Detection of Negative Content (Hoax) On Microblog DataThat Contains Covid-19 Information, *Syntax Literate: Jurnal Ilmiah Indonesia, 7* (6).
E-ISSN: 2548-1398
Published by: Ridwan Institute

monitoring the information circulating, clarifying the actual information, and taking action against hoax spreaders(Kemkominfo, 2021).

The purpose of this study is to develop a model that can detect hoaxes automatically in tweets on Twitter to reduce the spread level.

Study Literature

A. CRISP-DM

Cross-Industry Standard Process for Data Mining or CRISP-DM is a standard in data mining processing. CRISP-DM was built in 1966 to use it for data mining, analytics, and science projects(Sihombing, Jayadi, Chandra, & Liu, 2020). In CRISP-DM the data mining process is divided into 6 stages consisting of business understanding, data understanding, data preparation, modeling, evaluation and deployment as shown in Figure 1.

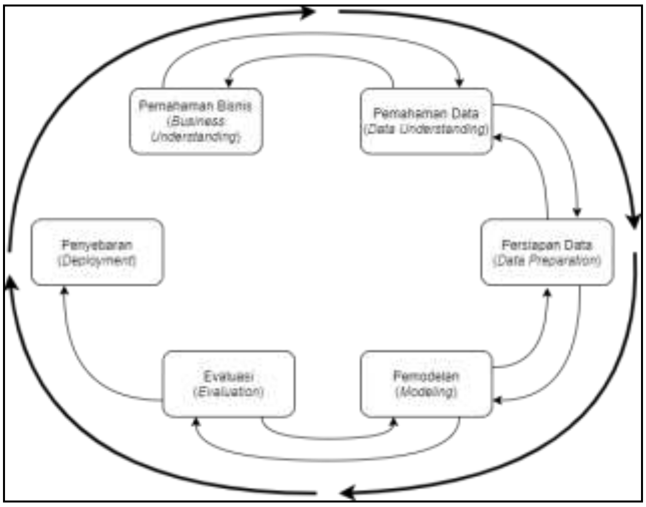
Pemahaman Bisnis (*Business Understanding*)
Pemahaman Data (*Data Understanding*)
Penyebaran (*Deployment*)
Persiapan Data (*Data Preparation*)
Evaluasi (*Evaluation*)
Pemodelan (*Modeling*)

Figure 1. Illustration of stages in CRISP-DM

B. DATA PREPARATION

Data Acquisition, Labeling, and Preprocessing
Step of preprocessing:

- *Case folding*, the stages of unfirming characters into lowercase letters.
- *Filtering*, steps to remove URLs, hashtags, hyperlinks and emoticons.
- *Tokenizing* is the process of dividing text into certain parts based on punctuation marks, numbers, words, and others.
- *Translate*, the process of translating words that contain repeated letters, so that the words “noo”, “no” and “Noooooo” are translated as the same word, namely “no”.
- *Stopword Removal*, removes words that have no meaning, such as “at”, “at”, “to”, “which”, and others

C. Synthetic Minority Over-sampling Technique (SMOTE)

The Synthetic Minority Over-Sampling Technique, or SMOTE, is used to overcome unbalanced data sets (Chawla, Bowyer, Hall, & Kegelmeyer, 2002; Pangastuti, 2018). Data is said to be unbalanced or imbalanced if the number between the classes is not equally represented(Chawla et al., 2002). The principle of applying

SMOTE so that a data set is balanced by increasing the number of samples in the minor class to be equivalent to the major class. Minor class addition generates synthesis data based on the nearest neighbor (k-nearest neighbor).
D. Classification Algorithm

Naïve Bayes

The Naïve Bayes classification method combines the supervised learning method (a learning method using sample data that has a label to then be used for new data classification) and probability classification(Parsian, 2015). The basic principle of nave Bayes is to apply Bayesian theory (from Bayesian statistics) with strong independent (naive) assumptions(Lesmana, 2013). In general, the formula for the Naïve Bayes algorithm is as follows(Walia, Rana, & Kansal, 2018):

$$P(H|A) = \frac{P(A|H)P(H)}{P(A)}$$

- P(H|A): Hypothesis Probability in data set A
- P(A|H): Probability of data set A in Hypothesis
- P(H): Probability of the hypothesis (prior probability)
- P(A): Probability of the observed sample data

Support Vector Machine (SVM)

The Support Vector Machine or SVM was first described by Vapnik, Bernhard Boser, and Isabelle Guyon in 1992(Ritonga & Purwaningsih, 2018). SVM is an algorithm that uses nonlinear mapping to convert the original training data to a higher. The model formed by SVM is in the form of a hyperlane, which is a function that is used to separate two different classes.

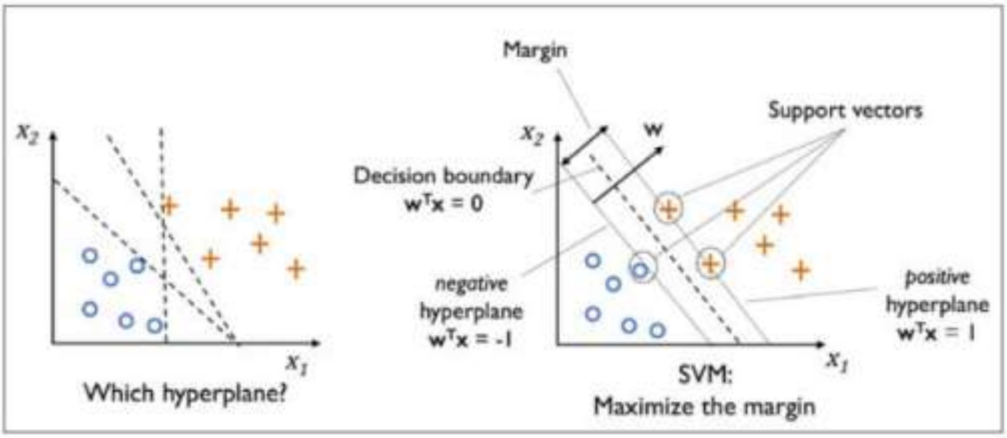

Figure 2 Illustration of support vector machine

Decision Tree

Decision tree is a decision-making method based on a flow diagram shaped like a tree consisting of several parts, namely the root node, internal node, and leaf node, as illustrated in *Figure 3*(Sá et al., 2016). The data obtained is included in the root node category, while the internal node contains statements from the data. A leaf node is a problem solving or decision-making.

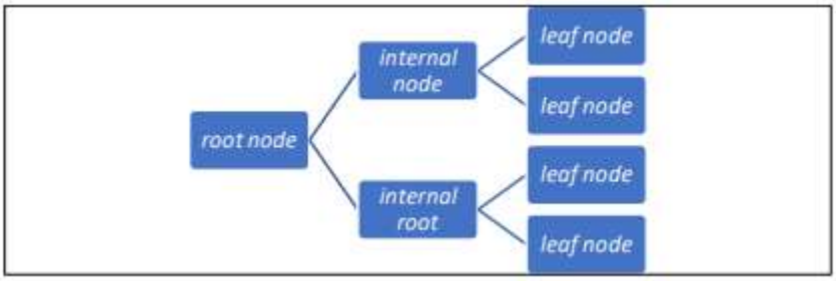

Figure 3 Illustration of decision tree components

Random Forest

Random forest was developed by Breiman in 2001 with the aim of improving the prediction process for the bagging method(Pangastuti, 2018). The random forest method is a collection of trees (decision tree) combined into a model, as shown in Figure 4 (Al-Ash, Putri, Mursanto, & Bustamam, 2019). From a collection of single trees, it is expected to have a small correlation result to obtain a smaller variety of estimates.

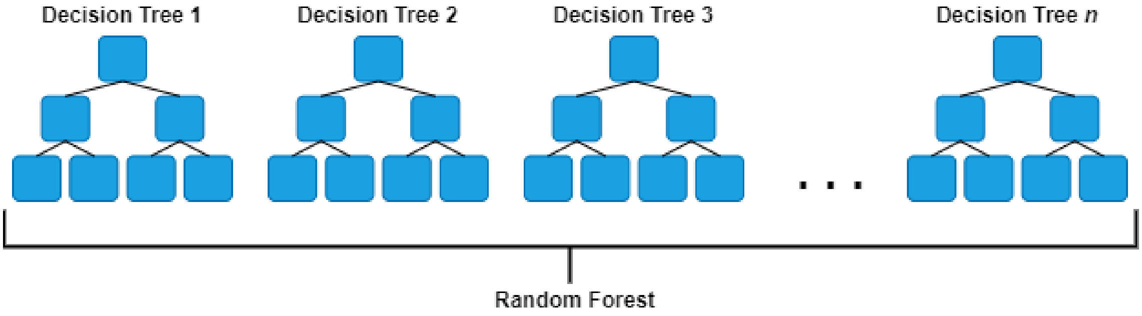

Figure 4 Illustration of random forest

E. Evaluation

K-Fold Cross Validation

K-Fold Cross Validation is a method to evaluate the performance of a model. The evaluation process is carried out by dividing the data into subsets of k partitions of the same size, where k indicates the number of partitions (Hulu, 2020). For each partition, a modeling process will be carried out with a performance test. Figure 5 illustrates the k-fold cross validation process where the data set is divided into 5 partitions. The 1st iteration shows that partition 1 is the test data and the other partitions are the training data. The 2nd iteration process is carried out like the process in the 1st iteration, but for the test data using partition 2 and the training data using a partition other than partition 2. Treat the replacement of test data and training data until the last partition.

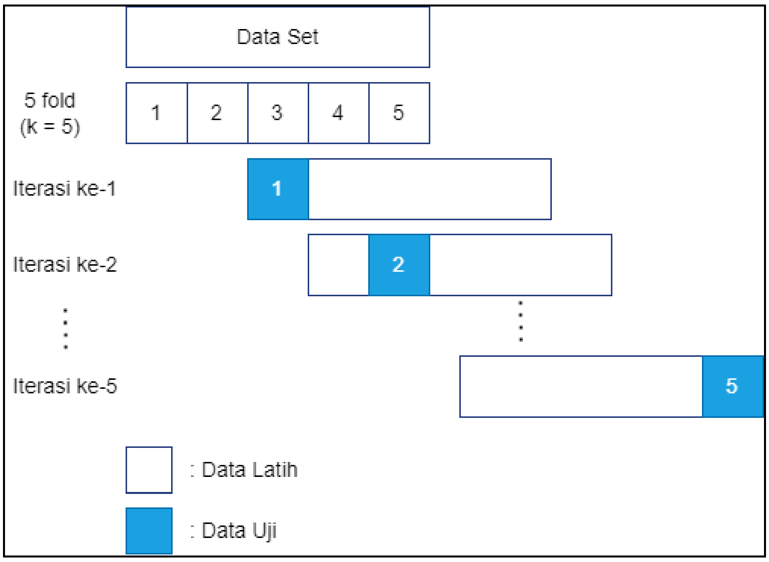

Figure 5 Illustration of K-Fold Cross Validation

Confusion Matrix

Confusion matrix merupakan tabel yang merangkum serta menunjukkan performa dari algoritma *machine learning* (Widaretna, Tirtawangsa, & Romadhony, 2021). Komponen penyusun *confusion matrix* terdiri atas *true positive* (TP), *true negative* (TN), *false positive* (FP) dan *false negative* (FN).

- TP: positive data predicted correctly
- TN: negative data predicted correctly
- FP: negative data predicted as positive data
- FN: negative data predicted as negative data.

The use of the confusion matrix is shown to provide information about the types of errors made by the prediction model. From the confusion matrix, there are several performances that can be known from a prediction model including accuracy, precision and recall.

Research methods

At this stage, a problem search is carried out using the Gap Analysis Technique to find common problems that exist. Problem identification is done with documents available on the internet and can be accessed by anyone. These documents are in the form of the Kominfo Strategic Plan, Kominfo Annual Report, Kominfo Performance Report and statistical data from survey results by several survey institutions. The data used comes from tweets containing information on COVID-19 or Corona on Twitter. The data collection process is carried out using a crawling technique with an API that has been provided by twitter.

Results and Discussion

From the labeling results, it can be seen that the comparison of the number of tweets between the non-hoax and hoax categories is quite far (1586 and 231) so that the sample data obtained is included in the unbalanced data category (imbalance). To overcome this, the SMOTE method was applied so that the hoax identification model obtained had better results.

The hoax identification model obtained is the result of machine learning using sample data that has been applied to the SMOTE method as learning data and test data. There are several algorithms used to create a hoax identification model, including Gausian Naïve Bayes, Multinomial Nave Bayes, Bernaulli Naïve Bayes, Decision Tree, Random Forest and Support Vector Machine.

In making the hoax identification model, the distribution of sample data into learning data and testing data was given two treatments. The first treatment was to share learning data and test data with a ratio of 70:30, hereinafter referred to as modeling without the application of k-fold. The second treatment was using the k-fold cross validation method with the configuration k=5.

To determine the best hoax identification model, a confusion matrix is used to measure the performance of each algorithm. Because the data used is included in the imbalance category, the f1-score value of each model reflects the performance of the model.

The results of the evaluation of the hoax identification model for each algorithm using the confusion matrix for the first treatment can be seen in Table 5.1 and Figure 5.1. From the evaluation results, the SVM algorithm shows the highest f1-score achievement with a value of 96 with precision, recall and accuracy values of 95, 98 and 99%

Table 1 Results of hoax identification modeling without applying K-Fold

Algoritma	*Precision*	*Recall*	*F1-Score*	*Accuracy*
Gaussian Naïve Bayes	87	100	93	98
Multinomial Naïve Bayes	85	100	92	98
Bernoullil Naïve Bayes	100	88	94	99
Support Vector Machine	95	98	96	99
Decision Tree	27	93	42	71
Random Forest	100	88	94	99

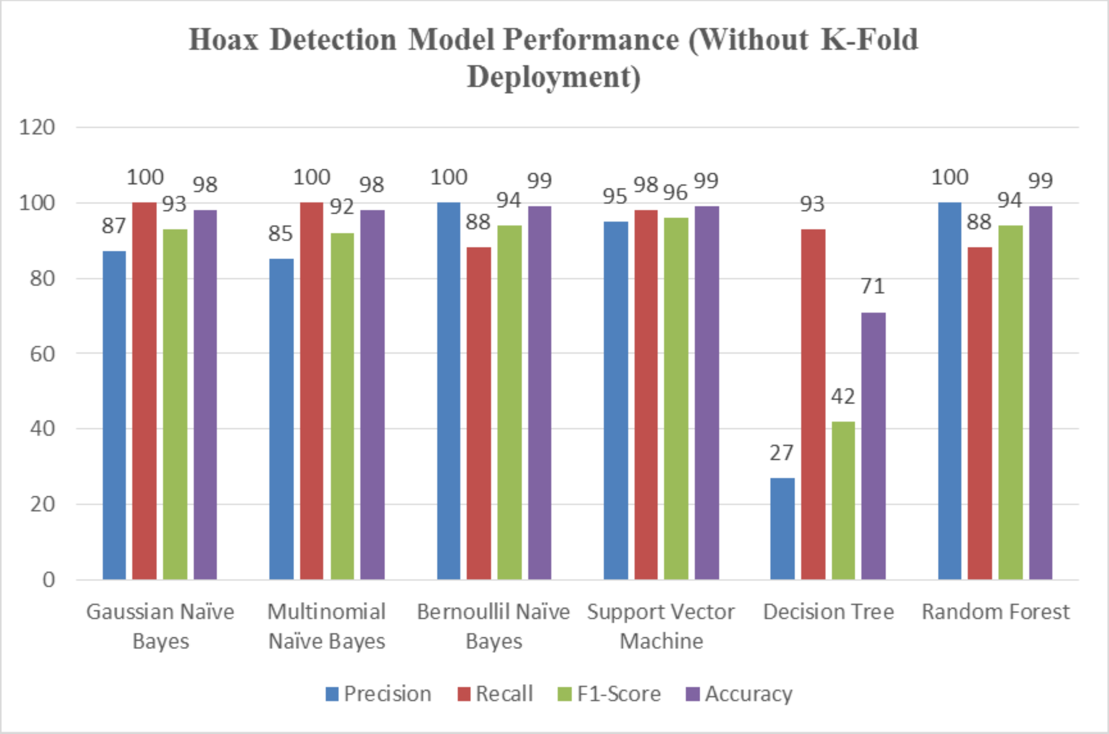

Figure 1 The performance of the hoax detection model without the application of k-fold

In The Second Treatment, The Results Of The Evaluation Of The Hoax Identification Model For Each Algorithm Can Be Seen In Table 5.2 And Figure 5.2. Based On The Evaluation Results In The Second Treatment, The Highest F1-Score Value Was Obtained By Modeling The Decision Tree Algorithm With A Value Of 83 With Precision, Recall And Accuracy Values Of 85.4, 81.4 And 97.2%.

Table 2 Results of Hoax Identification Modeling by Applying K-Fold

Algoritma	*Precision*	*Recall*	*F1-Score*	*Accuracy*
Gaussian Naïve Bayes	75,2	84	80,2	95,6
Multinomial Naïve Bayes	77,8	87,8	79,4	87
Bernoullil Naïve Bayes	79,2	57,4	64,8	87,8
Support Vector Machine	60	1,2	3	89
Decision Tree	85,4	81,4	83	97,2
Random Forest	80	71,2	75,4	96,6

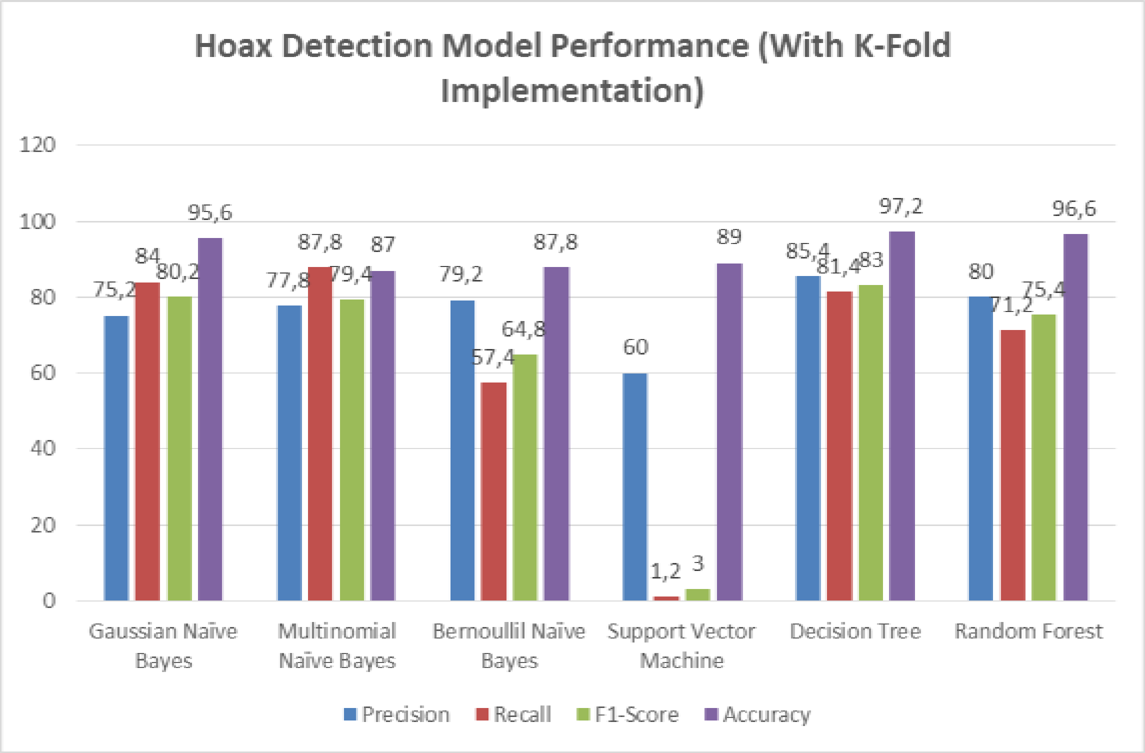

Figure 2 The performance of the hoax detection model with the application of k-fold

The best hoax identification model obtained is applied to all crawled data. The model used is the best model with the application of the k-fold method because it is more reliable. The results of the identification as shown in Figure 5.3, of the 18,170 tweets carried out by the hoax identification process, there were 10,104 tweets that were identified as not hoaxes and 8,066 tweets that were identified as hoaxes.

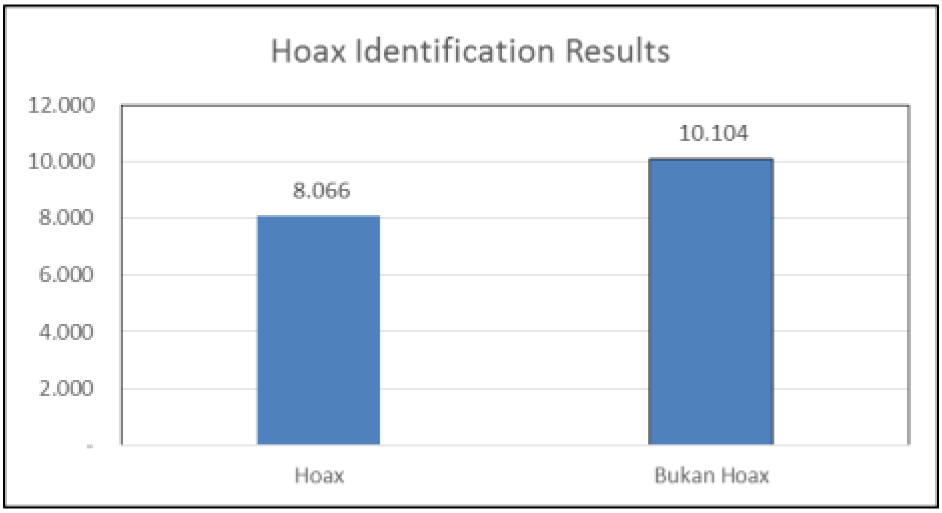

Figure 3 Hoax identification results

Tweets that are identified as hoaxes are carried out by a sentiment orientation classification process and generate a sentiment orientation as shown in Figure 5.4. From the results of the sentiment orientation classification, 3,820 tweets are classified as negative sentiment-oriented tweets and 4,246 tweets are classified as positive sentiment-oriented.

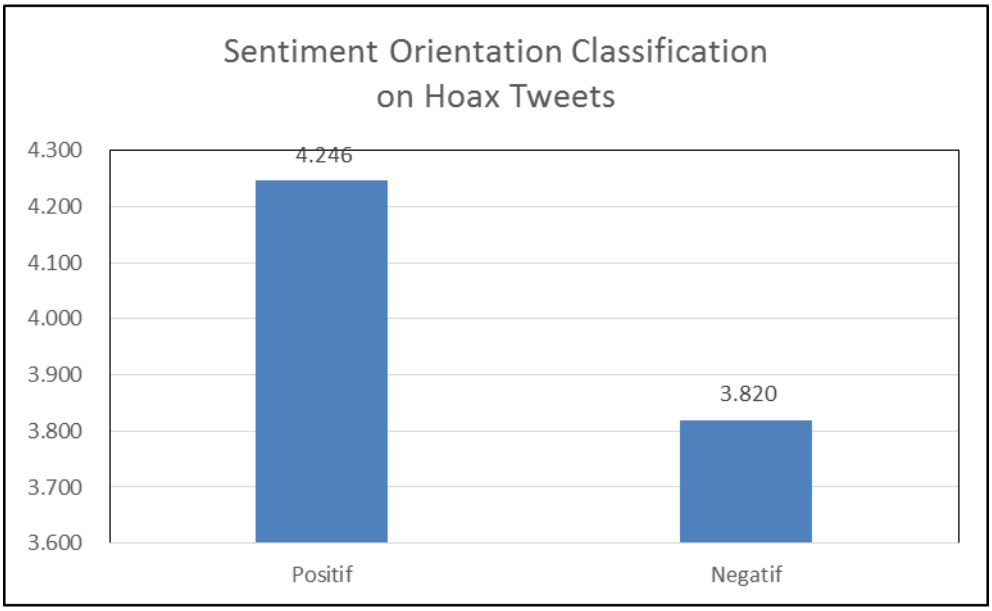

Figure 4
The results of the classification of sentiment orientation on twitter hoaxes

The author's expectations on tweets identified as hoaxes will be dominated by tweets with a negative orientation classification. This is because the characteristics of hoaxes are provocative and emotional.

Conclusion

The best hoax classification model for detecting potential hoaxes in tweets on Twitter uses the SVM algorithm for models without the application of k-fold with values of precision, recall, f1-score and accuracy of 95, 98, 96 and 99%.

The results of the classification of sentiment orientation on the classified hoax data using the best model (with the application of k-fold) obtained tweets with more positive sentiment orientation (52.64%). This is caused by several factors, namely the simple sentiment orientation classification method (using the lexicon) and the data used is less reliable because it uses data from the prediction model that is not 100% accurate.

BIBLIOGRAFI

Al-Ash, Herley Shaori, Putri, Mutia Fadhila, Mursanto, Petrus, & Bustamam, Alhadi. (2019). Ensemble Learning Approach on Indonesian Fake News Classification. *ICICOS 2019 - 3rd International Conference on Informatics and Computational Sciences: Accelerating Informatics and Computational Research for Smarter Society in The Era of Industry 4.0, Proceedings*. https://doi.org/10.1109/ICICoS48119.2019.8982409

Alamsyah, Syahdan. (2020). Heboh Isu Pasien Suspect Corona di Sukabumi , RS Belum Buka Suara.

Chawla, Nitesh V., Bowyer, Kevin W., Hall, Lawrence O., & Kegelmeyer, W. Philip. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*(Sept. 28), 321–357.

Hulu, Sitefanus(Universitas Sumatera Utara). (2020). Analisis Kinerja Metode Cross Validation Dan K-Nearest Neighbor Dalam Klasifikasi Data. In *Universitas Sumatera Utara*. Medan.

Juditha, Christiany. (2019). Literasi Informasi Melawan Hoaks Bidang Kesehatan di Komunitas Online. *Jurnal ILMU KOMUNIKASI*, *16*(1), 77. https://doi.org/10.24002/jik.v16i1.1857

Kemkominfo. (2021). *Laporan Tahunan 2020 (Indonesia Tekoneksi: Semakin Digital, Semakin Maju)*. Jakarta.

Lesmana, Pekik Indra. (2013). *Analisis Sentimen Pengguna Layanan Media Sosial Twitter di Indonesia*. Jakarta.

Pangastuti, Sinta Septi. (2018). *Perbandingan Metode Ensemble Random Forest Dengan Smote-Boosting Dan Smote-Bagging Pada Klasifikasi Data Mining Untuk Kelas Imbalance a Comparison of the Ensemble Random Forest Methods With Smote-Boosting and Smote-Bagging on Data Mining Classification Fo*. Surabaya.

Parsian, Mahmoud. (2015). *Data Algorithms: Recipes for Scaling Up with Hadoop and Spark* (1st ed.). O’Reilly Media, Inc.

Ritonga, Alven Safik, & Purwaningsih, Endah Supeni. (2018). Penerapan Metode Support Vector Machine (SVM) Dalam Klasifikasi Kualitas Pengelasan Smaw (Shield Metal Arc Welding). *Ilmiah Edutic*, *5*(1), 17–25.

Sá, J. A. S., Almeida, A. C., Rocha, B. R. P., Mota, M. A. S., Souza, J. R. S., & Dentel, L. M. (2016). *Lightning Forecast Using Data Mining Techniques On Hourly Evolution Of The Convective Available Potential Energy*. (March), 1–5. https://doi.org/10.21528/cbic2011-27.1

Sihombing, Pangondian Prederikus, Jayadi, Riyanto, Chandra, Edward, & Liu, Stefanie. (2020). Support vector machine-based hoax detection on indonesian online news. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(4), 6202–6207. https://doi.org/10.30534/ijatcse/2020/297942020

Somantri, Andri. (2020). Jangan Panik Corona! Warga Sukabumi Diminta Tak Usah Serbu Pasar.

Sutantohadi, Alief, & Rokhimatul Wakhidah. (2017). Bahaya Berita Hoax Dan Ujaran Kebencian Pada Media Sosial Terhadap Toleransi Bermasyarakat. *DIKEMAS (Jurnal Pengabdian Kepada Masyarakat)*, *1*(1), 1–5. https://doi.org/10.32486/jd.v1i1.153

Walia, Himdweep, Rana, Ajay, & Kansal, Vineet. (2018). A Naïve Bayes Approach for working on Gurmukhi Word Sense Disambiguation. *2017 6th International Conference on Reliability, Infocom Technologies and Optimization: Trends and Future Directions, ICRITO 2017*, *2018-Janua*, 432–435. https://doi.org/10.1109/ICRITO.2017.8342465

Wardani, Maria Magdalena Sinta. (2017). Manipulasi Bahasa dalam Teror Kabar Bohong (Hoax). *Sintesis*, *11*(2), 87–94.

Widaretna, Titi, Tirtawangsa, Jimmy, & Romadhony, Ade. (2021). Hoax Identification on Tweets in Indonesia Using Doc2Vec. *2021 9th International Conference on Information and Communication Technology, ICoICT 2021*, 456–461. https://doi.org/10.1109/ICoICT52021.2021.9527515

Copyright holder:
Putra Tresna Linge, Alfan Farizki Wicaksono (2022)

First publication right:
Syntax Literate: Jurnal Ilmiah Indonesia

This article is licensed under: