Journal of Engineering and Applied Technology
Vol. 4, No. 1, March 2023, pp. 40-51

Online ISSN 2716-2265
Printed ISSN 2716-2257

Comparative analysis of Indonesian news validity detection accuracy using machine learning

Rachelita Embun Safira [1], Akhsin Nurlayli [2,*]

[1,2]Universitas Negeri Yogyakarta, Indonesia
E-mail: akhsinnurlayli@uny.ac.id *
* Corresponding Author

ABSTRACT

Hoax news prediction is required to anticipate the growth of hoax news in social media. This study aimed to determine the best model for predicting whether the news is a hoax or valid based on the dataset taken from Kaggle.com. This study used several data prediction methods: Support Vector Machine (SVM), Random Forest, Logistic Regression, and Naïve Bayes. After the research processes and data testing, the results showed that the best model for predicting hoax news was SVM, which had the highest accuracy, precision, and recall score of the others.

ARTICLE INFO

Article history
Received:
23 February 2023
Revised:
20 March 2023
Accepted:
30 March 2023

Keywords
Logistic Regression
Naïve Bayes
Prediction
Random Forest
SVM

1. Introduction

The ease of accessing news today has become one of the advantages of technological advances. In this digital era, we can access news from printed media such as newspapers, tabloids, or magazines and from online news. The internet makes the news easiest to access in any media, such as websites, social media, and even any group chat. The speed of spreading news through internet access is one thing that triggers the emergence of new problems, namely the ease of spreading hoax news. The spread of untruth information can harm readers, which results in changes in human perception and thinking [1]–[3].

Hoax news is news or article whose credibility cannot be justified. People created hoax news to manipulate others, causing chaos, panic, and other bad purposes [4]. The worst impact of spreading hoax news is that it can trigger divisions and disputes between tribes, races, and countries [5]. From these problems, a system is needed that can detect the truth of news to reduce the impact of spreading hoax news. According to the Indonesian Ministry of Communication and Informatics, it is explained that a way to avoid hoax news is to always pay attention to the address of the news source sites [6]. However, many people need to get used to finding the news source before believing it.

Some studies indicated that machine learning methods were able to detect fake news. Some possible solutions to classify fake news in its current state have been identified and analyzed [7]–[9]. They comprehensively and systematically studied some research papers and projects on fake news detection. The results concluded about the streaming nature of fake news, and the computer was able to determine the false news by using advanced machine learning. Detecting fake news requires lifelong learning solutions due to the rapid changing of the content, style, language, and type of fake news. Some studies could have more clearly stated the machine learning-based fake detection system since numerous research only showed the detection performance without explaining the internal mechanics of the model process [10]. Integrating machine learning in semantic fake news detection has been studied by researchers [11],

[12]. Their experiments focused on short texts. They implemented deep learning and utilized the process in semantics features to improve accuracy. The accuracy increased up to 4.2%. The researchers concluded that due to the rapid growth of fake news, detection and the accuracy achieved would be depended on the datasets and the case study, a good model should be adaptive and not require much dataset fine-tuning. Also, the larger dataset would help to improve the model [12]. Some properties in the news to distinguish fake content have been explored by researchers. They trained a combination of machine learning algorithms using various ensemble methods and evaluated the performance on four real-world datasets. The experimental evaluation confirms their proposed ensemble learner approach's superior performance compared to individual learners [13]. Choosing inappropriate feature selection, inefficient tuning of parameters, and imbalanced datasets forced the poor accuracy of the fake detection model. An ensemble model to increase the accuracy compared to the existing models is needed. The ensemble model consisted of machine learning methods, i.e., Decision Tree, Random Forest, and Extra Tree Classifier. The results achieved a training accuracy of 99.8% and a testing accuracy of 44.15% for the ISOT dataset. Furthermore, a training and testing accuracy of 100% for the Liar dataset was achieved [14]. A systematic survey about fake news detection using deep learning and machine learning methods has been investigated. The analysis concluded that Naive Bayes, Support Vector Machine, Logistic Regression, and Random Forest were the most used in developing models for fake detection. However, each method had constraint(s) for a specific case. Since each case has its characteristics, it is recommended that the most appropriate method was trying and evaluating each [15]. A Random Forest for classifying and detecting fake news to enhance the accuracy of existing fake detection models with several machine learning methods was applied and ISOT News Dataset containing 44,919 news items was used. The dataset included 23,502 fake news and 2147 real news. The model utilized feature selection techniques such as chi, univariate, features importance, and information gain were proposed and achieved outstanding results in accuracy [16].

Regarding the previous works above, some machine learning methods effectively detect fake news, whether gained higher or lower accuracy for specific cases. Hence, this study aimed to compare the accuracy of models using the Support Vector Machine (SVM), Random Forest, Naïve Bayes, and Logistic Regression methods to predict the validity of the Indonesian news dataset and obtain a method which has a higher accuracy score for the Indonesian news validity detection model.

2. Theoretical Basis

2.1. Case Folding

Case folding is a method that can be used to equalize all the characters in data [17]. Case folding needs to be done to make the dataset easier to classify. Case folding is used in preprocessing, the first step in clustering and predicting data. In the dataset in the form of a collection of text, preprocessing is done by cleansing. One way of cleansing is case folding.

This study used case folding to convert data into lowercase or change all letters to lowercase. In addition to converting letters into lowercase and capital letters, case folding can also be used to remove some punctuation marks in the text [1].

2.2. Tokenizing

Tokenizing is a process that is used to break some text into some token. A token is a collection of characters separated by spaces or punctuation marks [18]. In tokenizing, punctuation marks and characters other than the alphabet will be automatically deleted because they are considered elements that

do not affect the sentence. In some tokenizing processing, the deletion of characters does not include in some numbers.

2.3. Stopwords Removal

Stopwords removal is one of the methods used in the data cleaning step for data text. Stopwords are words that do not have a significant impact on the meaning of a sentence. Natural Language Toolkit (NLTK) is one of the packages that can be used for the stopwords removal step. In NLTK, there are at least 758 stopword words in Indonesian [18]. In this package, there are various stopwords in various languages, including Arabic, French, German, Indonesian, English, and so on. Examples of stopwords in Indonesian are “yang”, “di”, “ke”, “dari”, andnd others. The use of stopword removal can be used to reduce the storage of words that are not important in memory. Using the stopword removal step, the process of checking the dataset will be lighter.

2.4. Term Frequency

Term Frequency (TF) is a method used to calculate the frequency of occurrence of a word in a dataset. Term Frequency is used to see the number of times a word appears. The number of the result of the term frequency process will show how important a word is in the document. If the frequency of a word is high, it indicates that the word is quite influential in a document and vice versa.

2.5. Machine Learning Algorithm

There are several algorithms that can be used for data classification and prediction, including:

a. Random Forest

Random Forest is an algorithm that can be used to classify large datasets. Random Forest uses a decision tree algorithm in its process. One of the advantages of using random forest is being able to classify data that has incomplete attributes. However, the drawback of this random forest is that it is not suitable for solving problems related to regression.

$$\text{Entropy } (Y) = \sum_i -P\left(\frac{C}{Y}\right) log_2 P(\frac{C}{Y}) \quad (1)$$

where, Y: Number of cases and $(\frac{C}{Y})$: Comparison of the value of Y to class C

b. Naïve Bayes

Naïve Bayes is a classification and prediction algorithm that is often used. This Naïve Bayes algorithm uses Bayes’ theorem. Naïve Bayes is a classification method based on simple probabilities to explain the assumption that there is no dependency relationship between variables.

$$P\left(\frac{X}{Y}\right) = \frac{P\left(\frac{Y}{X}\right)P(X)}{P(Y)} \quad (2)$$

where, $P\left(\frac{X}{Y}\right)$ is a probability of X over condition Y, $P\left(\frac{Y}{X}\right)$ is a probability of Y over hypothesis, $P(X)$ is probability X and $P(Y)$is probability Y. The advantage of this approach is that the classification will get a smaller error value when the data set is large. In addition, the Naïve Bayes classification is proven to have high accuracy and speed when applied to large databases [19].

c. Support Vector Machine (SVM)

SVM is one of the machine learning algorithms that can be used to make predictions in both classification and regression. It is a machine learning algorithm that implements a learning bias derived from static learning trained with a learning algorithm.

The SVM algorithm is used to find the best hyperplane by maximizing the distance between classes. Hyperplane is a function that can be used to separate between classes [17].

d. Logistic Regression

Logistic Regression is a model used to predict data in the form of binary, i.e. data is true false, true or false, or 0 1. TO predict the output variable that has a possible value of more than 2, the suitable algorithm used is multinomial logistic regression [20]. The logistic regression algorithm works by measuring the relationship between the target variable (the variable to be predicted) and the input variable or several features that will be used. The probability will be calculated based on the sigmoid function. The sigmoid function is a function that can change the values into the form 0 or 1.

2.6. Prediction

Dependent data prediction can be done by using machine learning which is built using a certain algorithm. Model development is done by training some data so that it can make decisions automatically. In this study, predictions were made to determine the appropriate category for news content. Furthermore, category data and news data will be trained to be able to make decisions about whether the news data is a hoax or real news. The data used in this research is a dataset obtained from Kaggle.com.

2.7. Accuracy

Calculating the prediction accuracy of the machine learning model that has been made is done by dividing the number of correct predictions by the number of predicted documents [1]. Accuracy calculations are needed to determine the level of accuracy of the model that has been made.

$$Accuracy = \frac{t_p + t_n}{t_p + f_p + f_n + t_n} \quad (3)$$

where t_p is the number of true positive data, t_n is the number of true negative data, f_p is for false positives, and f_n is for false negatives.

2.8. Precision

Precision is an indicator to see the comparison of the number of correct positive predictions with the total number of positive predictions. Precision is obtained by dividing the number of correct predictions t_p by the number of positive predictions obtained ($t_p + f_p$).

$$Prediction = \frac{t_p}{t_p + f_p} \quad (4)$$

Prediction is used to see the level of correct positive predictive data from a positive data set (true and false).

2.9. Recall

Recall is the ratio of correct positive predictive data to all correct data (true positive and false negative). A recall is commonly referred to as the sensitivity of machine learning.

$$Recall = \frac{t_p}{t_p + f_n} \quad (5)$$

3. Method

3.1 Data Source

The data used in this study were taken from 2 datasets from Kaggle, namely the dataset for hoax news and the dataset for valid news. These two datasets are combined for further training on the hoax or valid news prediction model. The dataset consists of 1000 hoax datasets and 1137 valid news datasets. So, the total dataset used in this study is 2137 news datasets.

Table 1. Dataset Attribute Description

Attribute	*Description*
ID	Is the unique id of each data in the dataset
Label	Contains dependent data which states that the news is valid or hoax news. 1 for hoax news, and 0 for valid news.
Judul	Contains the title of news
Narasi	Contains content of the news
Kategori	Conatins several categories according to news. Some data still have nan values in this attribute, therefore it needs to be processed first.

The dataset consists of hoax news and valid news stated in label attributes. Hoax news is marked with label '1'/ true, while valid news is marked with the label '0'/false. Details of the number of each news, both valid news and hoax news in the dataset are depicted in Fig. 1.

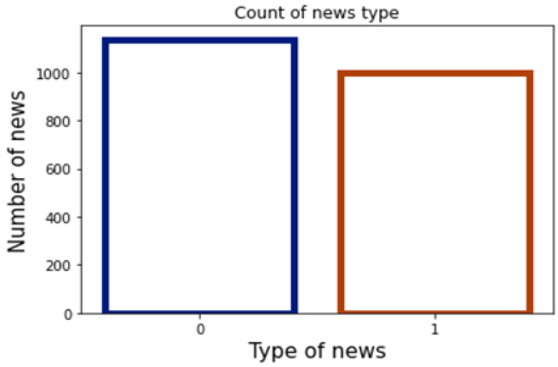


Fig. 1. Number of news for 2 types of news

In the dataset, news data is categorized into several categories according to the content of the news. However, there are still some news items that have a NaN value or are empty in the category attribute. Details of the amount of data in each category in this dataset are presented in the form of a diagram in Fig. 2.

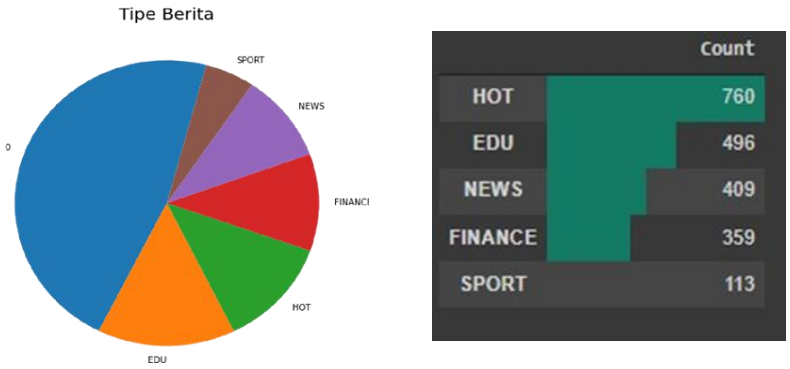


Fig. 2. Types and number of News

3.2 Model Creation Workflow

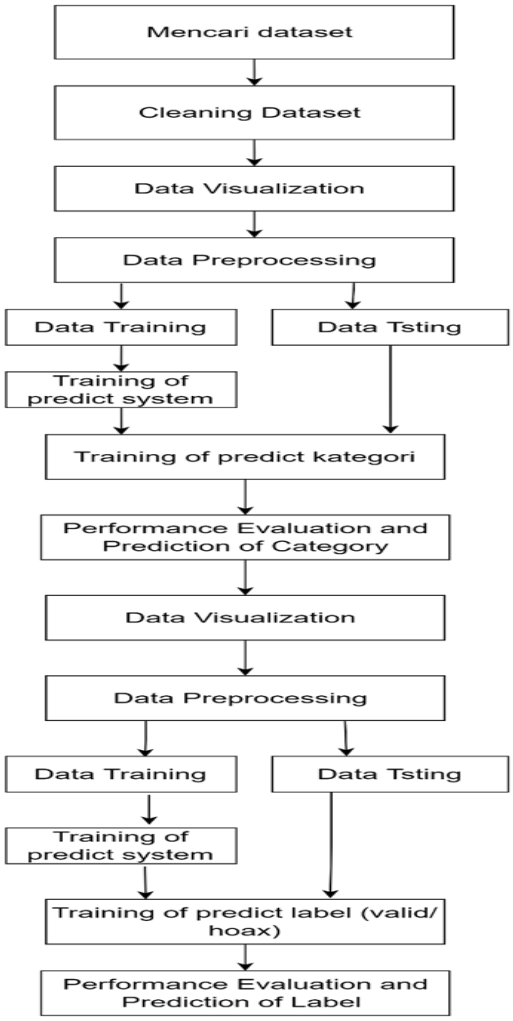


Fig. 3. Model Workflow

Fig. 3 is a design flowchart for the development of the prediction system through various stages to achieve the best result. The model developed is a model for predicting the category of news, and predicting labels whether the news is valid news or hoax news.

3.3 Data Source Collection

The first step in making this model is to collect relevant datasets to be used as test data for modelling. The dataset used in making this prediction model consists of a hoax news dataset and a valid dataset. The two datasets are combined into one dataset for model testing. Before being combined, the two datasets were cleaned to remove some unneeded attribute data.

3.4 Cleaning Dataset

The dataset cleaning phase is carried out to make the dataset clean and ready for preparation in the prediction model. The cleaning process is carried out by changing the data to lowercase letters to facilitate data checking. The cleaning process is also carried out to remove all stopwords on the site. This is done so that the substance of the news can be calculated appropriately. Next, is the stemmer process, which is a process to change all the words in the news into basic word forms. It is carried out using a literary stemmer package.

The next preprocessing step is the data tokenizer process using the RegexTokenizer library. The last step od text preprocessing is to calculate the value of each word using TF-IDF weighting. This process is carried out using the sklearn.feature library package, namely TfidfVectorier. In addition to using TF-IDF, this model also uses the CountVectorizer package which has a similar function to TfidVectorizer. The TfidfVectorizer package is used for data that contains many repetitions of the same word in one dataset. Whereas CountVectorizer is used for simple data, which does not contain many repetitions od the same word in one dataset.

The next process is to divide the data into 2 parts, namely train data (70 %) and test data (30%) to be implemented in the news category prediction model. The selection of train data dan test data is done randomly by the program, Train data and test data are taken from all data that has categorical attribute values (not NaN).

3.5 Category Data Prediction Model

The development of this category prediction model uses a library that is already available in sklearn called Logistic_Regression. The train data will be processed in Logistic_Regression. Furthermore, the model that has been trained will be tested using test data to see the accuracy score, recall score, and precision score for consideration in choosing a prediction model. News data with category values in the form of NaN is then processed using the data model to see the possible categories for the data.

3.6 Hoax News Data Prediction Model

Category data, news data, and news titles that have been carried out in the preprocessing step will then be processed to make a prediction model for the validity of the news. The model is made using SVM, Random Forest, Naïve Bayes, and Logistic Regression methods.

The model uses train data (70%) and test data (30%) from the results of category prediction data selected randomly by the program. From the results of the model, several methods will be compared with the accuracy of the predictions of the validity of the news. Despite this, the readers can find out which model has the highest level of accuracy in predicting hoax news data.

4. Results and Discussion

4.1. Dataset

Based on the two datasets that have been combined, the visualization of the merged datasets is as shown in Fig. 4.

a. ‘narasi’ attribute data

b. ‘judul’ attribute data

c. valid news data

d. hoax news data

Fig. 4. Visualization of data

From the visualization of the number of words that often appear, we can understand some of the words that are the main topic of each piece of news. The prediction model is carried out using categorical data with the following amounts as shown in Fig. 5.

	1.0	4.0	3.0	5.0	2.0
Kategori	0.281761	0.220126	0.202516	0.184906	0.110692

Fig. 5. Presentation of Category in Dataset

All categories of data values as shown in Fig. 6 are changed from string to float. The EDU category was changed to 1.0, SPORT into 2.0, HOT into 3.0, FINANCE into 4.0, and NEWS into 5.0.

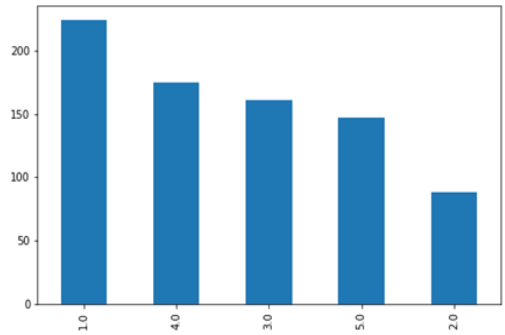


Fig. 6. Number of categories chart

The category prediction model was carried out using logistic regression and Naïve Bayes methods. The comparison of the accuracy values of the two methods is presented in Fig.7.

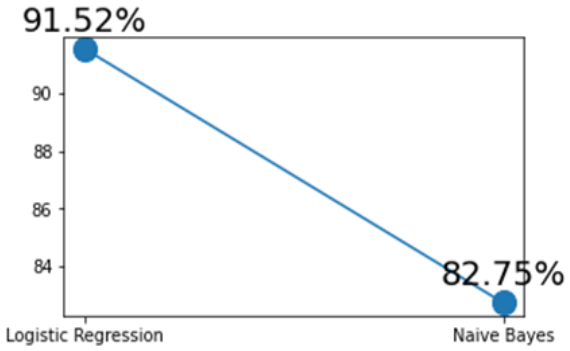
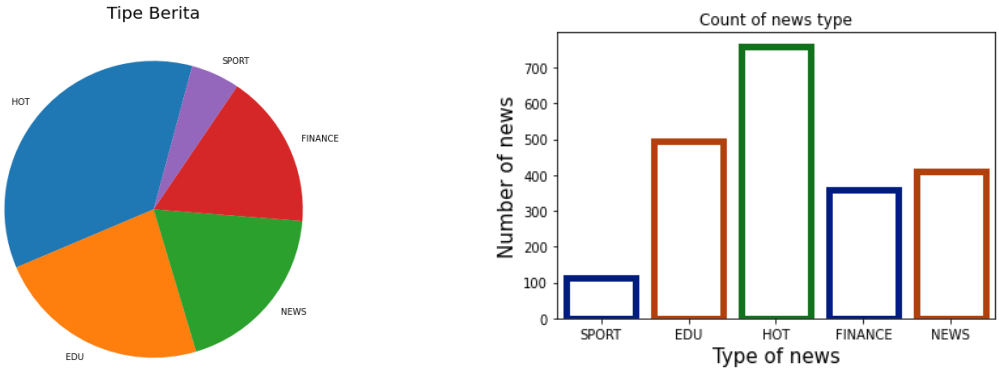


Fig. 7. Comparative diagram of Naïve Bayed and Logistic Regression models

From the calculation, the accuracy of training data from the Naïve Bayes and Logistic Regression methods, it was found that the accuracy value of the model with logistic regression was higher (91,52%) than using the Naïve Bayes (82.57%). Based on the consideration of the accuracy value, the model used to predict news categories is a model with the logistic regression method. Furthermore, data with the NaN category will be tested using a logistic regression model. The results of the category prediction test are presented in Fig. 8.



a. category of news b. number of news

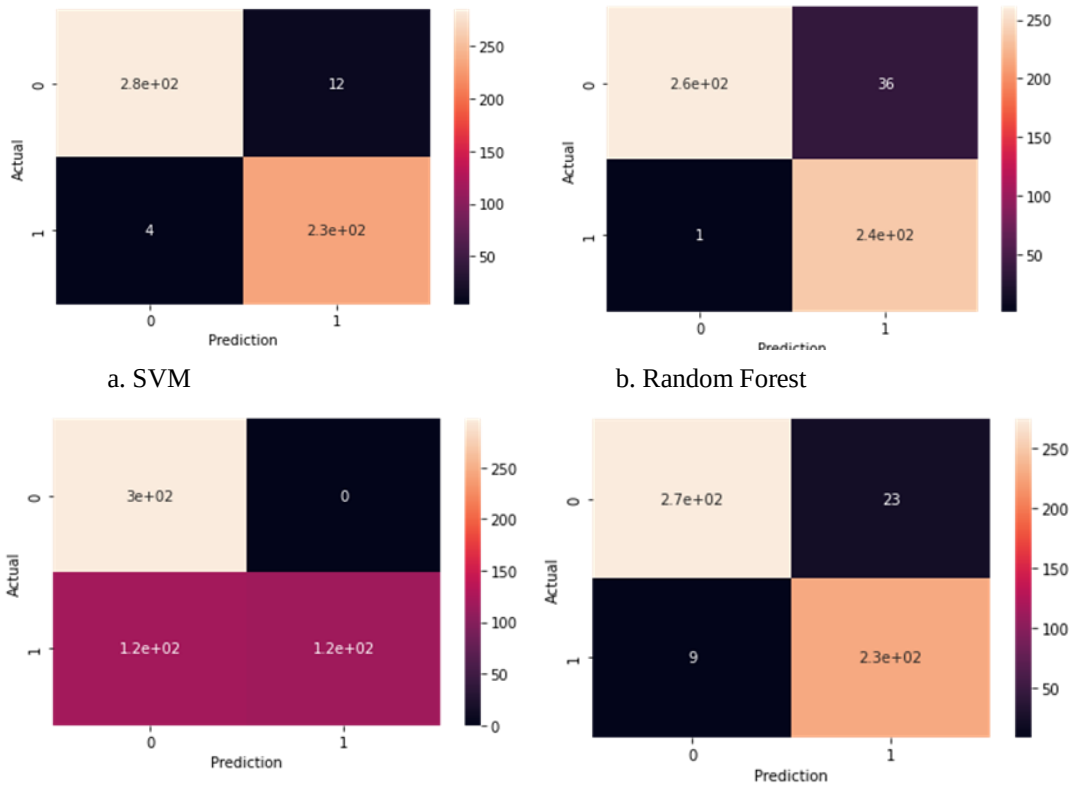
Fig. 8. Category and number of news diagrams after prediction

4.2. Hoax News Prediction Model

After the news data category has been successfully identified, then the data will be used for testing the hoax news prediction model. News data will be divided into 2 data, training data (70%) and testing data (30%) randomly. Several variables can affect the prediction decision of whether the news is hoax news or valid news. The "Clear Narasi" variable is a "narasi" variable that has been cleaned using a remove stopwords and stemmer process. In addition to the clean "narasi" variable, that affects decision-making is the "kategori berita" variable. Both of these variables become determinants of decision-making. All the data that has been processed is used to train a model for predicting hoax news. The model is to be trained with several methods. The results of data training and testing using these models can be seen in Table.2.

Table 2. The results of data testing using SVM, Random Forest, Naïve Bayes, and Logistic Regression Models

Model	*Accuracy*	*Recall*	*Precision*	*FI*
SVM	98.50%	98.50%	98.51%	98.50%
Random Forest	94.77%	94.77%	95.32%	94.78%
Naïve Bayes	85.05%	85.05%	88.03%	84.45 %
Logistic Regression	97.38%	97.38%	97.43%	97.38 %



a. SVM b. Random Forest

c. Naïve Bayes d. Logistic Regression

Fig. 9. Prediction heatmap for SVM, Random Forest, Naïve Bayes, and Logistic Regression Models

A comparison of the modelling results of each method is presented in the form of diagrams to make it easier to observe, as shown in Fig. 9. Based on the results of the accuracy test, it can be seen in Fig.10. that the model has the best accuracy value of around 98.50%. While the model with the lowest accuracy value is the Naïve Bayes model with an accuracy value of 85.05%.

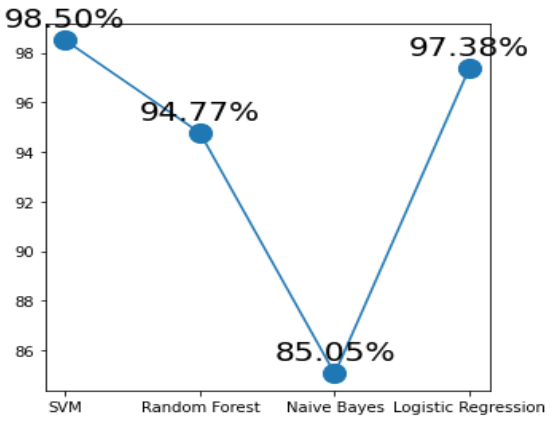


Fig. 10. Comparison of prediction method accuracy

5. Conclusion

Based on the results of the research and testing, the conclusion drawn includes the implementation of several methods to predict hoax news, including the SVM, Naïve Bayes, Logistic Regression, and Random Forest methods. The most optimum results or the model that produces the highest accuracy, precision score and recall values using the logistic regression method to create a category prediction model. Whereas in the hoax prediction model, the most optimum score was generated using the SVM method in predicting hoax news. Some of the things that support increasing the accuracy of predictions are the preprocessing process, including the data cleaning process with stopwords removal, and then the steamer does it. Furthermore, the data is divided into two, data train and data test. Both data are tokenized. Next, the data is processed with Tfidf Vector to calculate the occurrence of words in the news.

References

[1] F. Rahutomo, "Eksperimen Naive Bayes Pada Deteksi Berita Hoax Berbahasa Indonesia," *J. Penelit. Komun. dan Opini Publik Vol.*, vol. 23, no. 1, pp. 1–15, 2019, doi: 10.33299/jpkop.23.1.1805.

[2] D. Murthy *et al.*, "Bots and political influence: A sociotechnical investigation of social network capital," *Int. J. Commun.*, vol. 10, no. June, pp. 4952–4971, 2016.

[3] C. Zhang, A. Gupta, C. Kauten, A. V. Deokar, and X. Qin, "Detecting fake news for reducing misinformation risks using analytics approaches," *Eur. J. Oper. Res.*, vol. 279, no. 3, pp. 1036–1052, 2019, doi: 10.1016/j.ejor.2019.06.022.

[4] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Inf. Process. Manag.*, vol. 57, no. 2, p. 102025, 2020, doi: 10.1016/j.ipm.2019.03.004.

[5] Soleman, "Pemanfaatan Metode Klasifikasi Naïve Bayes Untuk Pendeteksi Berita Hoax Pada Artikel Berbahasa Indonesia," *J. CoreIT*, vol. 7, no. 2, pp. 83–93, 2021, doi: 10.24014/coreit.v7i2.14290.

[6] Kominfo, "Menkominfo Imbau Ekosistem Proaktif Cegah Hoaks," 2019. https://www.kominfo.go.id/content/detail/16276/menkominfo-imbau-ekosistem-proaktif-cegah-hoaks/0/berita_satker (accessed Jan. 26, 2023).

[7] M. Choraś *et al.*, "Advanced Machine Learning techniques for fake news (online disinformation) detection: A systematic mapping study," *Appl. Soft Comput.*, vol. 101, p. 107050, 2021, doi: 10.1016/j.asoc.2020.107050.

[8] F. A. Ozbay and B. Alatas, "Fake news detection within online social media using supervised artificial intelligence algorithms," *Phys. A Stat. Mech. its Appl.*, vol. 540, p. 123174, 2020, doi: 10.1016/j.physa.2019.123174.

[9] M. Davoudi, M. R. Moosavi, and M. H. Sadreddini, "DSS: A hybrid deep model for fake news detection using propagation tree and stance network," *Expert Syst. Appl.*, vol. 198, no. May 2021, p. 116635, 2022, doi: 10.1016/j.eswa.2022.116635.

[10] A. Aljarbouh, "Detecting Fake News using Machine Learning : A Systematic Literature Review," *Psychol. Educ.*, vol. 58, no. January, pp. 1932–1939, 2021, doi: 10.17762/pae.v58i1.1046.

[11] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," *Proc. Assoc. Inf. Sci. Technol.*, vol. 52, no. 1, pp. 1–4, 2015, doi: 10.1002/pra2.2015.145052010082.

[12] A. M. B. P. and R. Andonie, "Integrating Machine Learning Techniques in Semantic Fake News Detection Adrian," *Neural Process. Lett.*, 2020, doi: 10.1007/s11063-020-10365-x.

[13] I. Ahmad, M. Yousaf, S. Yousaf, and M. O. Ahmad, "Fake News Detection Using Machine Learning Ensemble Methods," *Hindawi*, vol. 2020, pp. 1–11, 2020.

[14] S. Hakak, M. Alazab, S. Khan, and T. Reddy, "An ensemble machine learning approach through effective feature extraction to classify fake news," *Futur. Gener. Comput. Syst.*, vol. 117, pp. 47–58, 2021, doi: 10.1016/j.future.2020.11.022.

[15] R. Varma and P. Churi, "A systematic survey on deep learning and machine learning approaches of fake news detection in the pre- and post-COVID-19 pandemic," no. October, 2021, doi: 10.1108/IJICC-04-2021-0069.

[16] M. Fayaz, A. Khan, M. Bilal, and S. U. Khan, "Machine learning for fake news classification with optimal feature selection," *Soft Comput.*, no. May, 2022, doi: 10.1007/s00500-022-06773-x.

[17] F. S. Jumeilah, "Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian," vol. 1, no. 1, pp. 19–25, 2017.

[18] Y. Sari, "Pengenalan Natural Language Toolkit (NLTK) Bagian 1," no. September, pp. 1–5, 2019.

[19] H. Jiawei, M. Kamber, J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[20] A. K. Santoso, A. Noviriandini, A. Kurniasih, B. D. Wicaksono, and A. Nuryanto, "Klasifikasi Persepsi Pengguna Twitter Terhadap Kasus Covid-19 Menggunakan Metode Logistic Regression," *J. Inform. Kaputama*, vol. 5, no. 2, pp. 234–241, 2021.