Faktor Exacta

Vol. 16, No. 1, March 2023, pp. 57~69
eISSN: 2502-339X, pISSN: 1979-276X, DOI: 10.30998/faktorexacta.v16i1.16486

Comparison of Classification Algorithms for Predicting Indonesian Fake News using Balanced and Imbalanced Datasets

Sayidati Karima[1], Achmad Benny Mutiara[2]
[1,2]Department of Management Information Systems, Gunadarma University, Indonesia

Article Info

Article history:

Received Feb 12, 2023
Revised Mar 03, 2023
Accepted Mar 15, 2023

Keywords:

Fake news
Logistic regression
Naive bayes
Random forest
Support vector machine

ABSTRACT

The rapid spread of information and fake news using internet media can have a bad impact and harm certain parties. In this research, a comparison was made between Logistic Regression, Naïve Bayes, Random Forest and Support Vector Machine algorithms to predict hoax news specifically for Indonesia. The system design stage starts from dataset collection, data labeling, pre-processing, TF-IDF weighting, model classification to testing. The highest accuracy results for both the number of unbalanced datasets and the number of balanced datasets were obtained from SVM with a comparison of training data and test data of 80:20. The unbalanced dataset has 85.47% accuracy and 90% F1-score and the balanced dataset has 84.36% accuracy and 84.80% F1-score. In this research, the unbalanced dataset provides better accuracy and test results using the Support Vector Machine algorithm.

Corresponding Author:

Sayidati Karima,
Department of Management Information Systems,
Gunadarma University,
Jl. Margonda Raya No. 100, Pondok Cina, Depok, 16424
Email: sayidatikarima@gmail.com

1. INTRODUCTION

Information dissemination and fake news using internet media and social media are things that can harm certain parties and can even influence other people in a bad direction because not all of the news that is spread is valid. Hoax news is news that contains things that are not in accordance with the facts or information that is actually not true, but is made as if it were true. Besides being able to affect people's psychology, hoax news can damage the reputation of the person concerned. The impact of spreading hoax news can also be dangerous for certain parties if it is misused and can cause losses from various aspects, such as public panic or can even lead to disputes in social relations and so on.

Advances in increasingly sophisticated technology can be utilized to prevent this problem. Hoax news can be detected with the help of Machine Learning. There are several stages that are passed, including collecting hoax news datasets, pre-processing to improve data quality such as classifying the text per word in hoax news first to help categorize the news as hoax news or not. From several researches, there are those who have predicted hoax news, including those using the Decision Tree Algorithm, Logistic Regression, Naive Bayes, Random Forest, Support Vector Macchine and many more. Each of these algorithms has various accuracy results, some have good accuracy of fifty percent and above and also not.

From the research Poddar et al. the highest accuracy results were obtained from Logistic Regression (91.6%). If using Count Vectorizer. Meanwhile, if using TF-IDF Vectorizer, the highest accuracy is obtained from SVM (92.8%) [1]. From the research Gowthami et al. the accuracy results obtained are SVM (70%) and Random Forest (98%) [2]. While the research Dhar et al. obtained the accuracy results obtained, namely, Logistic Regression (94%), Random Forest (93%), Naïve Bayes (90%) and Decision Tree 89 %) [3]. Then,

from the research Krishna et al. the accuracy results obtained, namely, Random Forest (65.6%), Naïve Bayes (63.7%), SVM (63%), Logistic Regression (62.5%), and Decision Tree (60%) [4]. From the research of Willy, et al. Comparison of the Random Forest Classifier Algorithm, Support Vector Machine and Logistic Regression Classifier on High Dimension Problems using Indonesian was also carried out. The dataset used in this study amounted to 20.000 using balanced datasets. The accuracy results obtained are Support Vector Machine (99.78%), Random Forest (99.73%) and the last Logistic Regression (99.20%) [5]. From research by Amanda Tabitha, et al. Hoax detection in Indonesian news about COVID-19 is carried out using an unbalanced dataset, from the accuracy results, it can be seen that the random forest method with the application of feature engineering produces an accuracy rate of 96.05%. Then, logistic regression (92.09%), SVM (91.53%), Naive Bayes (90.96%) and (90.40%) [6]. From six previous researches, it was found that the comparison of the same algorithm can give different results for prediction of hoax news in English and in Indonesian. But, the datasets used are also different, some use balanced and unbalanced datasets.

Because each of these algorithms has quite diverse results and the dataset used is also diverse, the authors try to compare the four algorithms that have better accuracy results, including Logistic Regression Algorithm, Naive Bayes, Random Forest, and Support Vector Macchine with compare balanced and unbalanced datasets from the same data source. What is taken for this research is a dataset of Indonesian news to determine the best model that will be used later to build a system that is useful for predicting hoax news. From this hoax news prediction research, it is hoped that it can provide the appropriate output so that it can focus more on developing the next hoax news prediction in Indonesia using the best algorithm.

2. METHOD

This method describes the methods used in system design to predict Indonesian hoax news by comparing several classification algorithms including Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machine using balanced and unbalanced datasets.

This chapter describes the flow of research methods used in developing a hoax prediction system or fake news from online news media. The stages of the system can be seen in Figure 1 as follows:

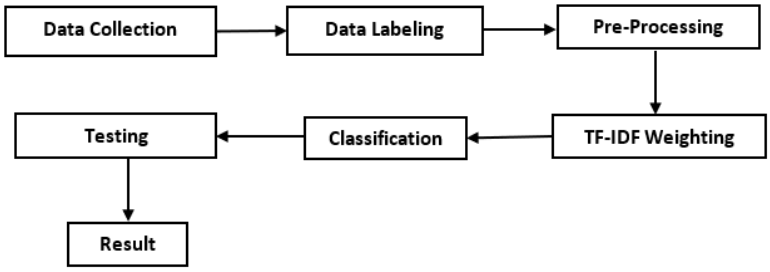

Figure 1. Stages of the System

2.1 Data Collection

The dataset taken is a collection of news containing narration and it is known that the news is hoax news or real news. The dataset is taken from the sites Github, Kaggle, Liputan6.com, turnbackhoax.id and Data Mendeley. The total dataset used is 8291 which contains 5797 hoax news and 2494 real news. Then, to test the results between the unbalanced datasets, a balanced dataset was also carried out with a total of 4988 datasets from the same data source which contained 2494 hoax news and 2494 real news.

2.2 Data Labeling

In the dataset taken there are several datasets that have not been labeled or also known as target attributes with a statement that the news is real or hoax. So that, at this stage data labeling will be carried out, real news will be labeled with a value of 0 which means "False" and hoax news will be labeled with a value of 1 which means "True".

2.3 Pre-Processing

Pre-processing of data is carried out to select text data so that it becomes more structured. In preparing the data, a pre-processing stage is needed. The purpose of this pre-processing stage is to improve the quality of the dataset used for training the hoax data model, such as cleaning words that are not really needed in the text. In the pre-processing process there are several stages that must be passed [16]. The pre-processing stage can be seen in Figure 2 as follows:

Figure 2. Pre-Processing Stage

2.4 Case Folding

The Case Folding stage is the initial part of the pre-processing stage. In the dataset obtained, not all letters are consistent with lower case or lower case. This Case Folding stage is needed in converting the entire text in the dataset into a standard form or into lowercase letters by using the lower() function to generalize text that is not yet in lowercase form [14]. The results of the case folding process are presented in Figure 3.

Before	After
Kominfo mengimbau masyarakat untuk tidak mempe...	kominfo mengimbau masyarakat untuk tidak mempe...
[BENAR] Dianggap Lecehkan Guru, Iklan Hago Dip...	[benar] dianggap lecehkan guru, iklan hago dip...
Jelas berbeda antara 'mengalirkan dana' dengan...	jelas berbeda antara 'mengalirkan dana' dengan...

Figure 3. Case Folding Stage

2.5 Data Cleansing

The Data Cleansing stage is a continuation of the previous stage which is no less important. This stage serves to eliminate characters that are less needed or not important, such as removing punctuation marks (commas, periods, exclamation points, and so on). Then, characters like (*), hashtag (#), username (@username) and other special characters. After that, the data is deleted using drop.duplicates if the same data is detected. The results of the data cleansing process are presented in Figure 4.

Before	After
kominfo mengimbau masyarakat untuk tidak mempe...	kominfo mengimbau masyarakat untuk tidak mempe...
[benar] dianggap lecehkan guru, iklan hago dip...	dianggap lecehkan guru iklan hago diproteskl...
jelas berbeda antara 'mengalirkan dana' dengan...	jelas berbeda antara mengalirkan dana dengan...

Figure 4. Data Cleansing Stage

2.6 Tokenizing

The tokenization stage is used to separate a sentence into word for word which is usually called a token so that it can be analyzed and facilitate data processing in the next stage [14]. The results of the tokenizing process are presented in Figure 5.

Before	After
kominfo mengimbau masyarakat untuk tidak mempe...	[kominfo, mengimbau, masyarakat, untuk, tidak,...
dianggap lecehkan guru iklan hago diproteskl...	[dianggap, lecehkan, guru, iklan, hago, diprot...
jelas berbeda antara mengalirkan dana dengan...	[jelas, berbeda, antara, mengalirkan, dana, de...

Figure 5. Tokenizing Stage

2.7 Filtering

The filtering stage is used to retrieve words that are less important from the token results by using a stoplist algorithm (removing less important words) or wordlist (saving important words). The goal is to remove commonly used words and words that do not have special meanings such as pronouns, prepositions, and conjunctions. This technique is also known as Stopwords Removal [17]. The results of the filtering process are presented in Figure 6.

 DOI: 10.30998/faktorexacta.v16i1.16486

Before	After
[kominfo, mengimbau, masyarakat, untuk, tidak,...	kominfo mengimbau masyarakat mempercayai situs...
[dianggap, lecehkan, guru, iklan, hago, diprot...	dianggap lecehkan guru iklan hago diproteskIar...
[jelas, berbeda, antara, mengalirkan, dana, de...	jelas berbeda mengalirkan dana memanfaatkan da...

Figure 6. Filtering Stage

2.8 Stemming

The stemming stage is done to change the word into its basic word form by removing the initial and final suffixes. The stemming process uses a special library for Indonesian language processing, namely the Python Sastrawi library [14]. The results of the stemming process are presented in Figure 7.

Before	After
kominfo mengimbau masyarakat mempercayai situs...	kominfo imbau masyarakat percaya situs situs m...
dianggap lecehkan guru iklan hago diproteskIar...	anggap leceh guru iklan hago diproteskIarifika...
jelas berbeda mengalirkan dana memanfaatkan da...	jelas beda alir dana manfaat dana temu kasus s...

Figure 7. Stemming Stage

2.9 TF-IDF Weighting

The next stage is weighting the value of the word using the TF-IDF (Term Frequency – Inverse Document Frequency) calculation. TF-IDF is a method of representing text data into numeric form. Term Frequency (TF) states the number of how many a term exists in one document. In contrast to TF, where the frequency of words appears, the higher the value. In Inverse Document Frequency (IDF), the less the number of times a word appears in the document, the higher the value [18]. The flow of the TF-IDF weighting stages can be seen in Figure 8.

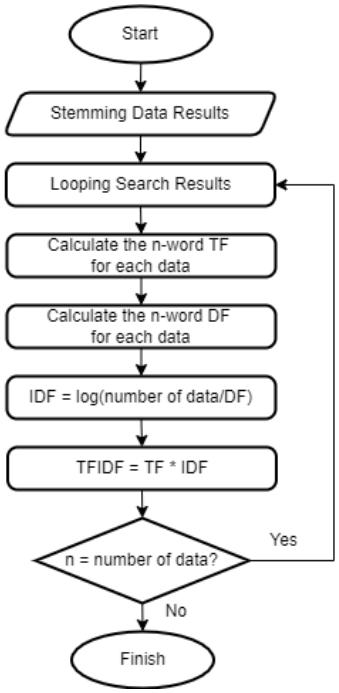

Figure 8. TF-IDF Weighting Flowchart

2.10 Model Classification

At this stage, the classification will be carried out using unbalanced datasets and balanced data. The model is classified using several algorithms, namely, Logistic Regression, Naïve Bayes, Random Forest and Support Vector Machine. The classification model with the best results will be used to test some news randomly. The classification model is a process used to classify test data to predict which news is hoax news or valid news. The steps are as follows:

1. Using the dataset that has been processed in the pre-processing stage until the value weighting uses TF-IDF so that it becomes a Document Term Matrix which can later produce a value of 0 as 'False' and a value of 1 as 'True'.
2. Then, each datasets with balanced and unbalanced data will be divided into two to be used at the training data stage with a comparison of 80% training data and 20% test data. The selection of training data and test data will be carried out randomly by the program.
3. Next, compare the results between algorithms based on the level of accuracy, precision, recall, and F-1 score of each algorithm.
4. After comparing the classification results, it can be determined which model will be used for testing by looking at the best accuracy results.

2.11 Implementation of The Model

In this implementation, it is the flow of stages of each model in classifying the dataset according to its category, which is fake news data or valid news data.

2.11.1 Implementation of Logistic Regression

At this stage, a classification model is made using Logistic Regression algorithm. The stages of Logistic Regression can be seen in Figure 9 as follows:

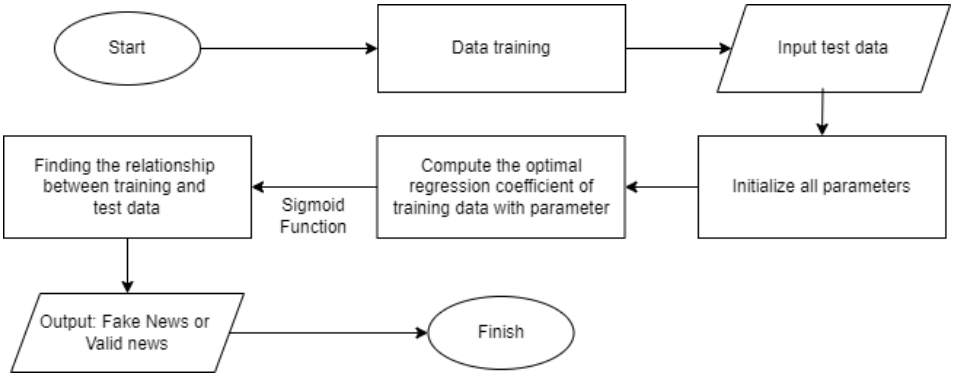

Figure 9. Logistic Regression Stages

The following is an explanation of Figure 9 which shows the stages of the Logistic Regression algorithm:

1. Load the dataset, and initialize all parameters needed such as predictor attributes, target attribute.
2. Compute the optimal regression of training data. In Logistic Regression to be able to get the theta value, the approach uses the sigmoid function and there are some calculation that must be done, namely calculate the predict dependent variable, cost function and calculate the gradient.
3. After all calculations are done, finding the relationship between test data and the probability of output results from training data.
4. If there is a relationship between the probability value and the data such as a value match, the target output will come out according to the match.

2.11.2 Implementation of Naïve Bayes

At this stage, a classification model is made using Naïve Bayes algorithm. The stages of Naïve Bayes can be seen in Figure 10 as follows:

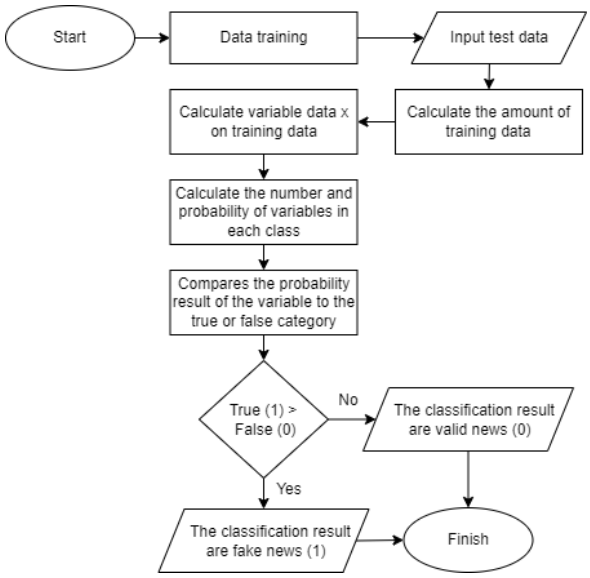

Figure 10. Naïve Bayes Stages

The following is an explanation of Figure 10 which shows the stages of the Naive Bayes algorithm:

1. Calculate the amount of training data. This calculation is carried out on testing the data.
2. Calculate variable data x on training data. This calculation aims to determine the total variables of each class.
3. Next, Calculate the number and probability of variables in each class. This probability calculation is using the gauss density equation in each class.
4. After calculating, the probability results will determine the probability of entering the category of true value or false value.
5. Then, the data is determined based on the probability results, the output will be obtained that the data is a category of fake news data or valid news data.

2.11.3 Implementation of Random Forest

At this stage, a classification model is made using Random Forest algorithm. The stages of Random Forest can be seen in Figure 11 as follows:

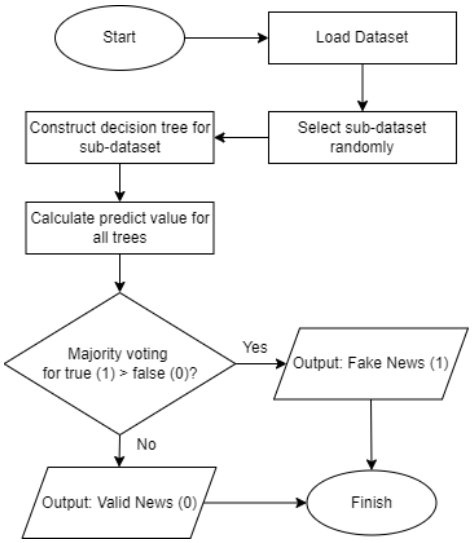

Figure 11. Random Forest Stages

The following is an explanation of Figure 11 which shows the stages of the Random Forest algorithm:

1. The algorithm selects a random sample from the provided dataset.
2. Make a decision tree for each selected sample. Then, the prediction results will be obtained from each decision tree that has been made.
3. The voting process is carried out for each prediction result. For this classification problem, use the mode value (the value that occurs most often).
4. The algorithm will choose the most voted prediction result (most votes) as the final prediction.

2.11.4 Implementation of Support Vector Machine

At this stage, a classification model is made using Support Vector Machine algorithm. The stages can be seen in Figure 12 as follows:

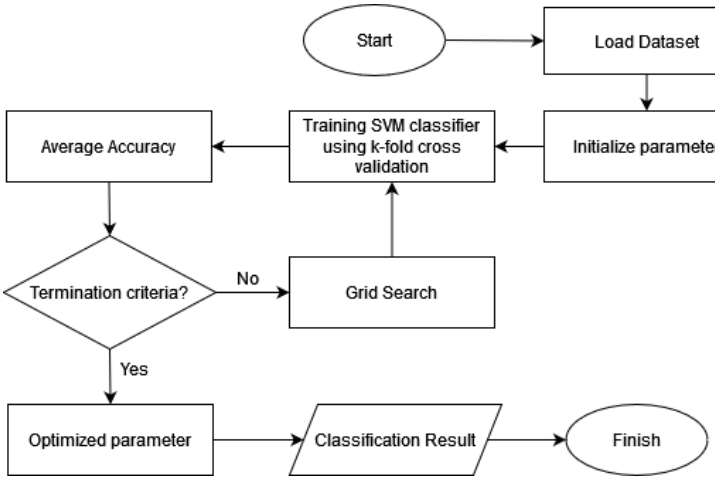

Figure 12. Support Vector Machine Stages

The following is an explanation of Figure 12 which shows the stages of the Support Vector Machine algorithm:

1. Initialize the parameter. Determining the value of the parameter C, has an effect on getting optimal parameters and can produce a hyperplane (separation line) to separate the two classes more optimally.
2. Training SVM classifier using k-fold cross validation. Cross validation serves to evaluate the performance of the algorithm where the data is separated into two subsets, namely learning process data and validation data.
3. Then, average the accuracy. If the termination meets the criteria, the results will be optimized. if not, the parameters will be determined again using a grid search to get the best results.
4. After selecting the best result, the classification will be obtained.

3. RESULTS AND DISCUSSION

This result shows the results of the comparison of the four algorithms that use a balanced dataset and unbalanced dataset to determine the best classification model to be used as a test in the next stage. The dataset taken is a collection of news containing narration and it is known that the news is hoax news or real news.

3.1 Implementation of Data Collection

The data training used was obtained from datasets taken from the Github, Kaggle, Liputan6.com, turnbackhoax.id and Mendeley Data sites, then all data was merged using a tool called DataMiner. The total dataset used is 8291 which contains 5797 hoax news and 2494 real news. Then, to test the results between unbalanced datasets, balanced datasets were also carried out with a total of 4988 datasets from the same data source which contained 2494 hoax news and 2494 real news.

Some of the news datasets are still not in the form of csv files or are still raw, so it is necessary to use tools for scrapping, such as the dataset obtained from the Liputan6.com and TurnBackHoax sites. An example of a page display that has been scrapped using the help of Data Miner tools can be seen in Figure 13 as follows:

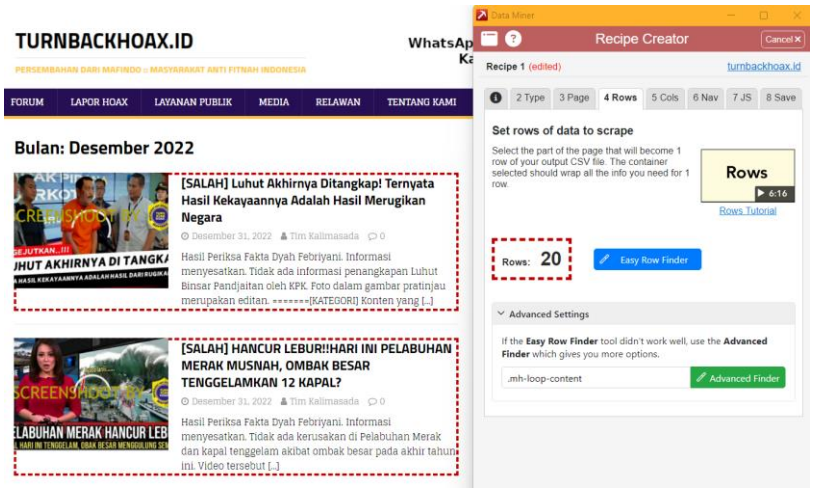

Figure 13. Example of A Scraped Page using the DataMiner

Then, when the scrapping is successful, the next step is to combine all the datasets obtained from various sources and then label the data. Datasets that have hoax news will be labeled 1 which means (True) and valid news will be labeled 0 which means (False).

3.2 Implementation of Word-Weighting

After the dataset obtained has completed the stemming stage, the next process is weighting the value of the word. Before weighting the words, the training data and test data were split first by comparing the train data (80) and test data (20). Comparison of the amount for balanced data is 998 for test data and 3990 for train data. Meanwhile, the comparison for unbalanced data is 6632 for train data and 1659 for test data. Next, weighting the value per word is carried out from the split training data and test data.

The source code uses TfidfVectorizer() which functions to convert text into a numeric representation so that the machine can easily process the data.

In this research, the weighting used was TF-IDF, the parameter values in TfidfVectorizer() function obtained using a balanced dataset were (3990, 20744) from train data, which means that there were 3990 document lines and 20744 unique words and for test data obtained (998, 20744). Meanwhile, the unbalanced datasets are (6632, 26128) from train data and (1659, 26128) from test data.

3.3 Classification Results and Evaluation Model

At this stage the model classification is carried out using several model algorithms, namely Logistic Regression, Naïve Bayes, Random Forest and Support Vector Machine. The purpose of this model classification is to find out the best accuracy results from the four algorithms in processing news text datasets.

In the formation of the SVM model the parameters tested were linear kernels with constant C values namely 0.1, 1, 10, and 100 and for parameter gamma are not needed in linear kernel. The best parameter optimization result is a linear kernel with a constant c value of 1 for unbalanced data and balanced data. Meanwhile, for other algorithms such as Naïve Bayes which is Multinomial Naïve Bayes the alpha parameter or a hyperparameter controls the form of the model itself. Then, for Logistic Regression and Random Forest following the standard parameters of the default model. Parameter results from SVM with various constant values can be seen in Table 1 as follows:

Table 1. Constant parameter results on SVM

Dataset Type	Parameter C	Accuracy
Unbalanced Data	0.1	73.35%
	1	85.47%
	10	83.54%
	100	82.21%
Balanced Data	0.1	74.84%
	1	84.36%
	10	82.16%
	100	81.76%

The following will explain the results of the comparison of the four algorithms, namely Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM) along with a comparison of balanced and imbalanced datasets. The results can be seen in Table 2 as follows:

Table 2. Comparison Results Based on Its Algorithm and Dataset Type

Dataset Type	Algorithm	TP	TN	FP	FN	Precision	Recall	F1 Score	Accuracy
Unbalanced Data	LR	1112	280	50	217	83.7%	95.7%	89.3%	83.9%
	NB	1142	138	20	359	76.1%	98.3%	85.8%	77.15%
	RF	1120	283	42	214	84%	96.4%	89.7%	84.56%
	SVM	1079	339	83	158	87.2%	92.9%	90%	85.47%
Balanced Data	LR	408	411	106	73	84.8%	79.4%	82%	82.06%
	NB	395	422	119	62	86.4%	76.8%	81.4%	81.86%
	RF	441	392	73	92	82.7%	85.8%	84.2%	83.46%
	SVM	434	408	80	76	85.1%	84.4%	84.8%	84.36%

Here are the details of the results of Precision, Recall, F1-score of the four algorithms using balanced and imbalanced datasets. The results can be seen in Figure 14, 15, 16, 17 as follows:

Unbalance Dataset

	precision	recall	f1-score	support
0	0.85	0.56	0.68	497
1	0.84	0.96	0.89	1162
accuracy			0.84	1659
macro avg	0.84	0.76	0.78	1659
weighted avg	0.84	0.84	0.83	1659

Balance Dataset

	precision	recall	f1-score	support
0	0.79	0.85	0.82	484
1	0.85	0.79	0.82	514
accuracy			0.82	998
macro avg	0.82	0.82	0.82	998
weighted avg	0.82	0.82	0.82	998

Figure 14. Classification Report of Logistic Regression

Unbalance Dataset

	precision	recall	f1-score	support
0	0.87	0.28	0.42	497
1	0.76	0.98	0.86	1162
accuracy			0.77	1659
macro avg	0.82	0.63	0.64	1659
weighted avg	0.79	0.77	0.73	1659

Balance Dataset

	precision	recall	f1-score	support
0	0.78	0.87	0.82	484
1	0.86	0.77	0.81	514
accuracy			0.82	998
macro avg	0.82	0.82	0.82	998
weighted avg	0.82	0.82	0.82	998

Figure 15. Classification Report of Naïve Bayes

Unbalance Dataset

	precision	recall	f1-score	support
0	0.87	0.57	0.69	497
1	0.84	0.96	0.90	1162
accuracy			0.85	1659
macro avg	0.86	0.77	0.79	1659
weighted avg	0.85	0.85	0.83	1659

Balance Dataset

	precision	recall	f1-score	support
0	0.84	0.81	0.83	484
1	0.83	0.86	0.84	514
accuracy			0.83	998
macro avg	0.84	0.83	0.83	998
weighted avg	0.83	0.83	0.83	998

Figure 16. Classification Report of Random Forest

Unbalance Dataset

	precision	recall	f1-score	support
0	0.80	0.68	0.74	497
1	0.87	0.93	0.90	1162
accuracy			0.85	1659
macro avg	0.84	0.81	0.82	1659
weighted avg	0.85	0.85	0.85	1659

Balance Dataset

	precision	recall	f1-score	support
0	0.84	0.84	0.84	484
1	0.85	0.84	0.85	514
accuracy			0.84	998
macro avg	0.84	0.84	0.84	998
weighted avg	0.84	0.84	0.84	998

Figure 17. Classification Report of Support Vector Machine

Based on the results of the comparison and testing between the Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machine models, there are many factors that determine the results, one of which is the dataset used and the number of datasets used also greatly affects the results. When viewed from the accuracy results, the highest accuracy value is in the Support Vector Machine, which is 85.47% with an unbalanced dataset type. However, when viewed from a balanced dataset, the SVM algorithm also obtains higher results, namely 84.36%. In addition to looking at the results of accuracy, it is also necessary to consider the results of precision, recall and F1-Score.

In this case the main target is to predict hoax news so that the more visible result is a target worth 1 (Hoax News) on precision and recall. Because, precision means what percentage of the data is correct compared to what is predicted to be correct. Meanwhile, recall calculates from the correct data how many percent are successfully predicted correctly. The value of precision and recall will affect F1-Score because F1-Score is a weighting between the two. So it is necessary to also consider the results of the F1-Score, and based on the results, the highest result of the F1 Score is obtained by the Support Vector Machine algorithm, for unbalanced dataset F1-Score is 90%.

Because, in certain cases if the False Negative is higher then it will be very dangerous. The meaning of False Negative is when the actual data is hoax news, but it is predicted that the news is real news will have a more bad impact. So, in the next test, this hoax news prediction system will use the Support Vector Machine model because it has better results.

3.4 Testing using Support Vector Machine

At this stage the model classification is carried out using several model algorithms, namely Logistic Regression, Naïve Bayes, Random Forest and Support Vector Machine. The purpose of this model classification is to find out the best accuracy results from the four algorithms in processing news text datasets.

In system testing, manual testing will be carried out on 20 data taken at random to be compared with the results of the test data and also compared using balanced datasets and imbalanced datasets. The data taken already has a label indicating that the news is hoax news (1) or real news (0). The results of the manual hoax news prediction test can be seen in Table 3 as follows.

Table 3. The Prediction Results on Random Data

No	News	Actual	Imbalanced Datasets	Balanced Datasets
1	Cerita Dalam Novel Telah Memprediksi Adanya Pandemi COVID-19	1	1 (True)	1 (True)
2	Kementerian Kesehatan Ukraina: 57 Orang Meninggal, 169 Orang Terluka di Seluruh Ukraina Akibat Serangan Rusia	1	1 (True)	0 (False)
3	Kementerian ESDM Buka Peluang Penyesuaian Harga Gas Khusus Industri	0	0 (True)	0 (True)
4	Garam dan Air Kelapa Dapat Hilangkan Vaksin di Dalam Tubuh	1	1 (True)	1 (True)
5	Di jejaring sosial, banyak beredar informasi yang menyebut lele sebagai ikan paling jorok. Dalam sesuap daging ikan lele, terkandung 3000 sel kanker.	1	0 (False)	0 (False)
6	Buntut Permasalahan TKI, Indonesia Kirim Ratusan Prajurit ke Perbatasan Malaysia	1	1 (True)	0 (False)
7	Harga Tiket Pesawat Mahal, Gelar Promo hanya Solusi Jangka Pendek	0	0 (True)	0 (True)
8	Banjir Bandang Hancurkan Hampir Seluruh Negeri, Pakistan Minta Bantuan Dunia	0	0 (True)	1 (False)
9	Kemunculan Ketua DPR RI Setya Novanto dan Wakil Ketua DPR Fadli Zon dalam rangkaian kampanye Donald Trump menjadi perhatian publik.	0	0 (True)	0 (True)
10	Air Rebusan Pare Dapat Membersihkan Kolesterol dan Semua Penyakit	1	1 (True)	1 (True)
11	Kalangan peneliti memperkirakan tsunami 'raksasa' dengan tinggi gelombang 20 meter akan terjadi di wilayah Jawa. Masyarakat pun diminta untuk berhati-hati.	0	1 (False)	1 (False)
12	Aksi balapan liar di Kota Surabaya, Jawa Timur yang memakan korban jiwa, tak hanya tengah ramai diperbincangkan di Tanah Air.	0	0 (True)	0 (True)
13	Omicron Bukanlah Virus Melainkan Akibat Dari Keracunan Chemtrail Yang Disebar Di Udara	1	0 (False)	1 (True)

14	Gaji pokok presiden yang besarnya hanya sekitar Rp30 juta per bulan saat ini dianggap sangat kecil Pemerintah berencana mengusulkan kenaikan gaji presiden agar layak menjadi Rp553 juta tiap bulannya.	1	0 (False)	0 (False)
15	Sebuah Koran Menyebut Bill Gates Melakukan Depopulasi Melalui Pemaksaan Vaksinasi	1	1 (True)	1 (True)
16	Kim Seon Ho Donasi Rp 590 Juta untuk Yayasan Leukemia Anak	0	1 (False)	1 (False)
17	Terdapat Banyak Manfaat Dari Kunyit Putih Bagi Kesehatan, Mencegah Berbagai Jenis Kanker	0	1 (False)	1 (False)
18	Soal Opsi Kenaikan Harga BBM, Wapres: Kalau Subsidinya Ditambah, Membahayakan APBN	0	0 (True)	0 (True)
19	Minyak Makan Merah Mulai Distribusi Januari 2023, Harga Rp 9.000 per Liter	0	1 (False)	1 (False)
20	Bahaya Racun dari Nasi yang dimasak dengan Magicom	1	1 (True)	1 (True)

In Table 3 it is found that from the 20 data taken at random, for imbalanced dataset there are 7 data predictably wrong and for balanced dataset obtained 9 data predictably wrong. Based on the input test data, a lot of news data that contains numbers produces wrong predictions using both balanced datasets and unbalanced datasets. It can be seen that in this system there are still shortcomings in the detection of news texts containing numbers.

Meanwhile, the prediction results using imbalanced dataset from the test data amounting to 1659 data can be seen in Figure 18 as follows.

	Data Uji	Label Asli	Label Prediksi	Keterangan
1023	akun whatsapp bupati malang sanusi hasil perik...	1	1	Benar
3989	kim jong un kena kemeja putih tulis bersih rak...	1	1	Benar
1088	foto jalan tol trans papuahasil periksa fakta ...	1	1	Benar
357	jemaah calon haji minta jangan gantung jemur s...	0	1	Salah
189	buat yg aku sayang meng ingat kan kalau dlm wa...	1	1	Benar
...	...	...	...	...
73	rusuh jadi wilayah dekai kabupaten yakuhimo pa...	0	0	Benar
324	balai besar awas obat makan bbpom pasti produk...	0	0	Benar
3596	la gode serang asrama tni bukan serang asrama ...	1	1	Benar
173	pidato mark zuckerberg tsinghua university chi...	0	1	Salah
1452	info dr dinas sehat kota tangerang permen jari...	1	1	Benar

1659 rows × 4 columns

Figure 18. Prediction Results from Test Data using Imbalance Data

The total number of hoax news predictions from the Test Data using Imbalance Dataset, it was found that from the actual label and prediction label that had the same value (True) there were 1418 data. Meanwhile, the wrong data amounted to 241 data.

Whereas, the prediction results using balanced dataset from the test data amounting to 998 data can be seen in Figure 19 as follows.

	Data Uji	Label Asli	Label Prediksi	Keterangan
433	video heroik tni pecah kaca mobil pakai tangan...	0	1	Salah
23	jakarta pidana kasus bom thamrin dodi suridi h...	0	0	Benar
2704	pasien positif covid rst latumeten ambon berit...	0	0	Benar
2502	salah bakar duta perancis sudan tolak gambar n...	1	1	Benar
709	salah link cek bantu per september besar guna ...	1	1	Benar
...	...	...	...	...
314	benar kemendagri pernah keluar edar surat beka...	0	0	Benar
1159	salah video roket al qassam hantam wilayah isr...	1	1	Benar
70	jakarta perilaku masyarakat tusuk jarum silet ...	0	0	Benar
582	gubernur sumatera selatan alex noerdin kata ru...	0	0	Benar
3256	salah video shalat subuh jamaah semenjak hagia...	1	1	Benar

998 rows × 4 columns

Figure 19. Prediction Results from Test Data using Balance Data

The total number of hoax news predictions from the Test Data using Balance Dataset, it was found that from the actual label and prediction label that had the same value (True) there were 848 data. Meanwhile, the wrong data amounted to 150 data.

Based on the prediction of hoax news using data taken at random or from test data using an imbalance dataset and a balanced dataset, the correct results are more than the wrong ones. But when viewed from the reality when testing the system with random data or with test data, the results still have some shortcomings. The existence of a system error in predicting news has many factors including the possibility of a limited dataset so that the word weighting process becomes less than optimal.

4. CONCLUSION

Based on the results of the model classification and the tests carried out, it can be concluded that this Indonesian hoax news prediction system gets good results. The results of the comparison of the unbalanced dataset and the balanced dataset both show that the highest classification results are obtained using the Support Vector Machine Algorithm. The results of the unbalanced dataset have an accuracy of 85.47% and an F1 score of 90% with a comparison of training data and test data of 80: 20. Meanwhile, a balanced dataset has the highest accuracy of 84.36% and an F1 score of 84.8%.

From testing hoax news using balanced and unbalanced training data, both show the same algorithm ranking, where Support Vector Machine takes first place, Random Forest second, Logistic Regression third and finally Naive Bayes. However, when compared with the overall results obtained between SVM, Random Forest, Logistic Regression it is not too far away so it can be concluded that these three algorithms have good performance in the classification of hoax news and depend on the dataset and parameters used. While the results of Naive Bayes always occupy the last position and the difference is quite far using a balanced dataset.

In addition, unbalanced datasets with a number of main class targets where there is more fake news data get better results than balanced datasets. However, in some cases, such as news text containing numbers, the truth cannot be known accurately because in the TF-IDF the weighting of numbers does not affect the calculation so that it has no effect whatsoever. So for further research, it is better for this hoax detection system to use the Support Vector Machine as the main choice with more dataset comparisons for the main class targets and use more specific news categories so that they are not too broad in order to get more accurate results.

This research is useful for minimizing the spread of hoax news circulating in society. However, eliminating it completely can be said to be almost impossible, so the provision of education to the public also really needs to be improved to be able to understand the circulation of news that is not necessarily true.

REFERENCES

[1] K. Poddar, *et al* "Comparison of Various Machine Learning Models for Accurate Detection of Fake News," Innovations in Power and Advanced Computing Technologies, 2019, doi:10.1109/i-PACT44901.2019.8960044.
[2] K. Gowthami, *et al*, "Identification of Fake News through SVM and Random Forest," IJESC. Vol. 10 No. 10. ISSN 2321-3361, 2020.
[3] M. Dhar, *et al*, "Detection of Fake News using Machine Learning Algorithms," IJARIIE. Vol. 7 Issue 4. ISSN(O) 2395-4396, 2021, doi: 10.1109/ICAC3N53548.2021.9725560
[4] R. Krishna, *et al*, "Survey on Fake News Detection using Machine Learning Algorithms," International Journal of Engineering Research & Technology. ISSN 2278-0181, 2021, doi: 10.17577/IJERTCONV9IS08026.
[5] Willy, *et al*, "Perbandingan Algoritma Random Forest Classifier, Support Vector Machine dan Logistic Regression Clasifier Pada Masalah High Dimension (Studi Kasus: Klasifikasi Fake News)", Jurnal Media Informatika Budidarma, ISSN 2614-5278, Vol. 5, No. 4, Page 1720-1728, 2021, doi: 10.30865/mib.v5i4.3177.
[6] Amanda, T., et al, "Deteksi Hoaks Pada Berita Berbahasa Indonesia Seputar COVID-19", Jurnal Ilmiah Teknik Informatika Format. ISSN : 2089 – 5615, Vol. 10 Nomor 1, 2021, doi:10.22441/format.2021.v10.i1.007.
[7] A. Thoha, "Respon Mahasiswa Jurusan Komunikasi UIN Suska Riau Terhadap Program Siaran Suskapucino di Radio Suska FM 107,9 Mhz Pekanbaru," Indonesia, Universitas Islam Negeri Sultan Syarif Kasim Riau, 2018.
[8] C. Juditha, "Interaksi Komunikasi Hoax di Media Sosial serta Antisipasinya Hoax Communication Interactivity in Social Media and Anticipation," Jurnal Pekommas. Vol. 3 No.1:31-44, 2018, doi: 10.30818/jpkm.2018.2030104.
[9] A. Kadir, "Logika Pemrograman Python", Jakarta: PT. Elex Media Komputindo, 2019.
[10] H. S. Simon, "Penentuan Posisi Objek Berbasis Image Processing Dengan Menggunakan Metode Convolutional Neural Network," UIB: Universitas Internasional Batam, 2020.
[11] M. S. Shell, An Introduction to Numpy and Scipy. USCB Engineering, 2019.
[12] D. Dewanti, *et al*, Bootcamp Data Science (Machine Learning, Bogor : Inspira Pustaka Aksara, 2021.
[13] E. J. Rifano, *et al*, "Text Summarization Menggunakan Library Natural Language Toolkit (NLTK) Berbasis Pemrograman Python," ILKOMNIKA. Vol. 2 No.1:8-17. E-ISSN 2715-2731, 2020, doi: https://doi.org/10.28926/ilkomnika.v2i1.32
[14] S. Wahyunita, "Analisa Sentimen Tweet Berbahasa Indonesia Dengan Menggunakan Metode Pembobotan Hybrid TF-IDF Pada Topik Transportasi Online," UMM: Universitas Muhammadiyah Malang, 2018, doi:10.22219/repositor.v2i2.238.

[15] D. F. Setiawan, *et al*, "Aplikasi Web Scraping Deskripsi Produk," Jurnal TeknoInfo. Vol. 14 No. 1. ISSN: 2615-224X, 2020, doi: 10.33365/jti.v14i1.498.
[16] L. Hermawan and M. Bellaniar, "Pembelajaran Text Preprocessing berbasis Simulator Untuk Mata Kuliah Information Retrieval," Transformatika. Vol. 17 No. 2. ISSN: 1693-3656, 2020, doi: http://dx.doi.org/10.26623/transformatika.v17i2.1705
[17] A. Hakim, "Klasifikasi Sentimen Terhadap Bukalapak Dengan Menggunakan Metode Naïve Bayes Classifier," Fakultas Sains dan Teknologi. UIN Sultan Syarif Kasim Riau, 2018.
[18] A. N. Assidyk, E. B. Setiawan, and I. Kurniawan, "Analisis Perbandingan Pembobotan TF-IDF dan TF-RF pada Trending Topic di Twitter dengan Menggunakan Klasifikasi K-Nearest Neighbor," E-Proceeding of Engineering. Vol. 7 No. 2 Hal 7773. ISSN: 2355-9365, 2020, doi: https://doi.org/10.34818/eoe.v7i2.12794
[19] D. T. Wisudawati, "Analisis Sentimen Terhadap Dampak COVID-19 pada Performa E-Commerce di Indonesia Menggunakan Support Vector Machine (Review Aplikasi Tokopedia Pada Google Play)," Seminar Nasional VARIANSI, 2020. ISBN: 978-602-53397-2-1.
[20] H. F. Putro, *et al*, "Penerapan Metode Naive Bayes Untuk Klasifikasi Pelanggan," Jurnal TIKomSiN, Vol. 8, No. 2. ISSN: 2620-7532, 2020, doi:10.30646/tikomsin.v8i2.500.
[21] H. Nalatissifa, W. Gata, S. Diantika, K. Nisa, "Perbandingan Kinerja Algoritma Klasifikasi Naive Bayes, Support Vector Machine (SVM), dan Random Forest untuk Prediksi Ketidakhadiran di Tempat Kerja," Jurnal Informatika Universitas Pamulang. Vol. 5 No. 4, Hal. 578-584. ISSN: 2541-1004, 2020, doi: http://dx.doi.org/10.32493/informatika.v5i4.7575
[22] F. Ridzuan, *et al*, "A Review on Data Cleansing Methods for Big Data," Procedia Computer Science 161(3):731-738, 2019.