

Support Vector Machine For Hoax Detection

Ni Wayan Sumartini Saraswati¹, I Putu Krisna Suarendra Putra², I Dewa Made Krishna Muku³, Gede Dana Pramitha⁴

^{1,2,3,4} Teknik Informatika, Institut Bisnis dan Teknologi Indonesia

Jl. Tukad Pakerisan No.97, Panjer, Denpasar Selatan, Denpasar, Indonesia

e-mail: sumartini.saraswati@gmail.com¹, spkrisna2001@gmail.com², dewamuku@instiki.ac.id³,
danagede@instiki.ac.id⁴

Received : June, 2023

Accepted : July, 2023

Published : August, 2023

Abstract

Along with the development of information technology, news media has also developed by presenting information online. Along with the rapid development of online news, the spread of fake news information (hoaxes) is also increasing rapidly and widely. Hoax news is often spread intentionally for various purposes. Generally, hoax news aims to direct the reader's perception to believe in a bad perception of an event, character or even a company. The motivation is to invite readers to believe something that is not true with the aim of benefiting the news disseminator is something dangerous. This research aims to detect English-language hoaxes by applying the Support vector machine (SVM) algorithm. In this study, the data used are two data sources, namely English news datasets from Kaggle and English news taken from BBC. The results of this study show that the application of the SVM algorithm turns out to get good performance because the model is able to classify hoax news with an accuracy of 99.4% on Kaggle data while on the BBC news dataset the model gets an accuracy of 98.9%. This research also shows that the SVM method is proven to have good generalization properties. Where it is able to identify test data that is completely different from the training data.

Keywords: hoax news, SVM, text classification, BBC news

Abstrak

Seiring dengan perkembangan teknologi informasi, media berita juga ikut berkembang dengan menyajikan informasi secara online. Bersama dengan perkembangan berita online yang begitu pesat, maka penyebaran informasi berita bohong (hoax) juga semakin cepat dan meluas. Berita hoax seringkali disebarkan dengan sengaja untuk berbagai macam tujuan. Umumnya berita hoax ini bertujuan mengarahkan persepsi pembaca untuk mempercayai sebuah persepsi buruk mengenai sebuah kejadian, tokoh atau bahkan perusahaan. Motivasi yang dilakukan adalah mengajak pembaca mempercayai sesuatu hal yang tidak benar dengan tujuan agar menguntungkan pihak penyebar berita merupakan sesuatu hal yang berbahaya. Penelitian ini bertujuan untuk melakukan deteksi pada berita palsu (hoax) berbahasa Inggris dengan menerapkan algoritma Support vector machine (SVM). Dalam penelitian ini data yang digunakan berupa dua sumber data yaitu dataset berita berbahasa Inggris dari Kaggle dan berita berbahasa Inggris yang diambil dari BBC. Hasil dari penelitian ini menunjukkan penerapan algoritma SVM ternyata mendapatkan performa yang baik karena model mampu mengklasifikasikan berita hoax dengan akurasi 99,4% pada data Kaggle sedangkan pada dataset berita BBC model mencapai akurasi hingga 98,9%. Dalam penelitian ini, metode SVM terbukti memiliki sifat generalisasi yang baik. Dimana mampu mengidentifikasi data uji yang benar benar berbeda dari data latih.

Kata Kunci: *berita hoax, SVM, klasifikasi teks, berita BBC*

1. INTRODUCTION

Hoax or fake news is information engineered by irresponsible parties with the aim of spreading misleading opinions in the public sphere. According to some sources, a hoax is false information. Another source states that a hoax is information that is fabricated. The information is used to hide the actual of information. In addition, a fake news is an attempt to distort the facts. The facts will be replaced with information that is convincing but cannot be verified.

Hoaxes aim to create opinions, accompany public opinion, and shape readers' perceptions. There are various purposes for spreading hoaxes. Generally, it is used for for engaging in "black campaigns" to undermine competitors. In addition, the purpose of people spread fake news are obtain financial gain by increase site views, do manipulation of public opinion, create confusion, or individual entertainment [1]. Consequently, recipients are provoked to share them with their peers and make the hoaxes spread very fast.

This information will create unrest for many parties, such as news that can trigger divisions between individuals or groups. Misleading public opinion towards certain objects or parties will lead to confusion and misunderstanding of information circulating in the public sphere. Hoax news often targets people's emotions. The slander that is spread can ignite hatred and anger, so that people have a negative perspective on a person, group, or product. Because of the many hoaxes that have been circulating, it is difficult for people to distinguish between fake information and facts. With the spread of hoaxes, people no longer believe in the real facts because they are already mistaken [2].

The adverse effects of hoaxes can be very fatal. According to psychological analysis, there are several factors that influence a person to believe in fake news and hoaxes. The tendency factor is to deny what is happening. So that the individual concerned will believe anything that is contrary to the facts, the psychological tendency to believe in conspiracy theories, and ideological and political attachment to the spreader.

If hoax news can be detected early on, then the adverse effects caused by this news will be minimized. The public has begun to be smart and critical to understand and analyze all information in the public sphere, so that later there will be no more people who are harmed both psychologically and materially. Early detection of hoaxes will also break the network or chain of spread to the public. At least the public has realized from the beginning that the news is fake news with unclear sources.

The spread of fake news through various medium, especially online platform has not been stopped. The reason is there is no system that can control fake news with little or no human involvement. Experiments shows that machine learning algorithms have ability to detect fake news because there is initial set of cases to be trained on [3].

Detecting fake news is essential for a healthy society, and there are several different approaches to detecting fake news. From a machine learning point of view, fake news detection is one of binary classification cases [4].

Application of this machine learning can be applied in this research because machine learning is useful for solving problems in a scalable way and can process and analyze large and complex data in a shorter time. One example of the use of machine learning is to detect news whether it is fake news or not.

The problem of recognizing fake news is a problem that can be approached by solving text classification, so some machine learning methods that can be used to classify text can solve it. Research by [5] uses the decision tree method, to solve the detection of fake news in English.

Research conducted by [6] perform classification using two methods ensemble classifiers. Both methods consist of three learning models. The first ensemble is random forest, logistic regression, and KNN. The second are linear SVM, logistic regression, and CART (classification and regression trees).

Research by [7] compared several machine learning methods including XGBoost, Random Forest, Naïve Bayes, KNN, and SVM. While other studies compare Naïve Bayes, SVM, Neural Network and LSTM methods [8]. Seeing from some of these studies, SVM algorithm highly popular and great to go. SVM method offers high accuracy and works well in high-dimensional spaces. SVM classifiers basically use a subset of training points where to get the results this method uses very little memory so that computation becomes lighter. So, in this research we try to measure how SVM performs in solving the problem of fake news or hoax detection.

Some methods use the SVM method to solve fake news detection, as done by [9]–[11] showing that the SVM method is quite effective for solving fake news detection problems. However, it has not been done for the original news from the scrapping results so that SVM algorithm will use to solve the problem in this experiment. In addition, we also validate the model quality measurement using the 10 cross fold validation method.

2. RESEARCH METHOD

Text mining is one of methods to process text from documents (mining). The purpose of text mining is to identify words which may represent the content of the document, so it can analyze its relationship between document. In addition, text mining is used to extract the information that can be useful from the document or data source. This data sources have an unstructured format. So, by using text mining, data source will go through identification and exploration into interesting pattern. Text mining or text analytics convert that unstructured data into structured data [12]. The specific tasks of this methods include text categorization and text clustering. The commonly used method for text mining processing is the machine learning method. Text mining or text analytics

In general, it can be said that machine learning refers to an approach to data analytics carried out by automating the creation of analytics models. As a branch of AI, machine learning is about idea that computer can learn, find patterns, and make decisions from data, also with minimal human intervention in the process [13]. In machine learning, Support Vector

Machine (SVM) is one popular method which is used to perform text mining.

In machine learning, SVM is one of supervised learning to do classification and regression, such as Support Vector Classification and Support Vector Regression. The concept of SVM is more advanced than other classifications algorithm. In addition, SVM mathematically clearer. SVM also have linear and non-linear algorithm to solve classification and regression problems. An SVM is a binary linear non-probabilistic classifier since it takes a set of input data and predicts, for any given input, the likelihood that the input belongs to one of the two classes. An SVM training method creates a model that predicts whether incoming data will fall into one of the two categories given a training set that has been classified as belonging to one of the two categories.

In text classification, high dimensionality is one of the challenging problems that can be much more complicated and very expensive in terms of storage of data and execution time [14].

SVM algorithm greater efficiency when it used to higher dimensional spaces. In other side, SVM work effectively where the number of dimensional spaces exceeds the amount of sample values [15].

In a literal sense, the SVM model is a mapping of data as points in space, with categories of distinct examples being divided by as many distinct, broad gaps as feasible. Then, based on which side of the gap the data falls, new data is mapped into the same area and estimated to belong to a category. The SVM strategy makes use of mapping into a bigger space to make it simple to calculate the cross product in terms of variables in the origin space, resulting in an acceptable processing overhead. A kernel function $K(x, y)$ that can be selected based on the issue defines the cross product in the bigger space. The collection of points whose cross product with the vectors in a vast space is constant is referred to as a set of hyperplane in that space. One option for the vector defining the hyperplane is to linearly combine it with the parameter i from a database of feature vector images. With this selection, a feature space point x is mapped to a hyperplane that is described by the relation given below:

$$\sum a_i K(x_i, x) = \text{constant}_i \quad (1)$$

Each component of the measurement is a summation of the degree of proximity between the test point x and point x_i in the related database if $K(x, y)$ becomes small as y grows farther than x . The relative proximity of each test point to data points coming from either one of the clustering sets can then be determined using the aforementioned kernel sum.

Each component of the measurement is a summation of the degree of proximity between the test point x and point x_i in the related database if $K(x, y)$ becomes small as y grows farther than x . The relative proximity of each test point to data points coming from either one of the clustering sets can then be determined using the aforementioned kernel sum. [16]. The vector will divide the data into the best categorization form after finding the highest Hyperplane margin.

SVMs are considered as non-parametric learning algorithms since the number of parameters grows with the size of the training set. It is based on statistical learning theory, which decreases uncertainty in the structure of the model.

Text data in order to be processed using SVM, the data must first be converted into vector data. A feature extraction method is needed to characterize the data. The Term Frequency - Inverse Document Frequency (TF-IDF) method is a frequently used technique for feature extraction in text data.

According to [17] The TF-IDF algorithm approach is helpful for determining each word's weight in relation to how frequently it appears in the entire document. This approach is likewise renowned for being effective, simple, and producing precise outcomes. For each token (word) in each document in the corpus, this method will calculate the Term Frequency (TF) and Inverse Document Frequency (IDF) values. The TF-IDF approach, in short, is used to determine how frequently a word appears in a document. The mathematical notation for term frequency-inverse document frequency (TF-IDF) is as follows:

$$TF - IDF = TF \times IDF \quad (2)$$

TF-IDF is commonly used when converting text data into vectors but taking into account whether a word is informative enough or not. Using a collection of words, the TF-IDF statistical metric quantifies a term's significance to a document. In text mining, user modeling, and information retrieval, it is frequently employed as a weighting factor. The tf-idf value depends on how many documents in the corpus contain the phrase and rises according to the number of times it appears in the sentence.

A statistical technique called cross-validation (CV) can be used to assess how well a model or algorithm performs when the data is divided into two subsets: learning data and validation/evaluation data. The learning subset trains the model or algorithm, and the validation subset verifies it. The size of the dataset might also be taken into consideration while choosing a CV type. Because it can shorten computation times while preserving estimation accuracy, K-fold CV is frequently utilized.

For machine learning classification issues where the output can be two or more classes, the confusion matrix is a performance indicator. A table called the "Confusion Matrix" has four different sets of expected and actual values. In Confusion Matrix, the classification process produces four terms: True Positive, True Negative, False Positive, and False Negative.

Table 1: *Confusion Matrix*

	Actual Positive Class	Actual Negative Class
Predictive Positive Class	True Positive (tp)	False Negative (fn)
Predictive Negative Class	False Positive (fp)	True Negative (tn)

To calculate the performance of the classification model can be calculated using some measurement value.

Accuracy can be calculated by the formula:

$$Accuracy = \frac{tp+tn}{tp+fp+tn+fn} \times 100\% \quad (3)$$

Precision can be calculated by the formula:

$$Precision = \frac{tp}{tp+fp} \quad (4)$$

Recall can be calculated by the formula:

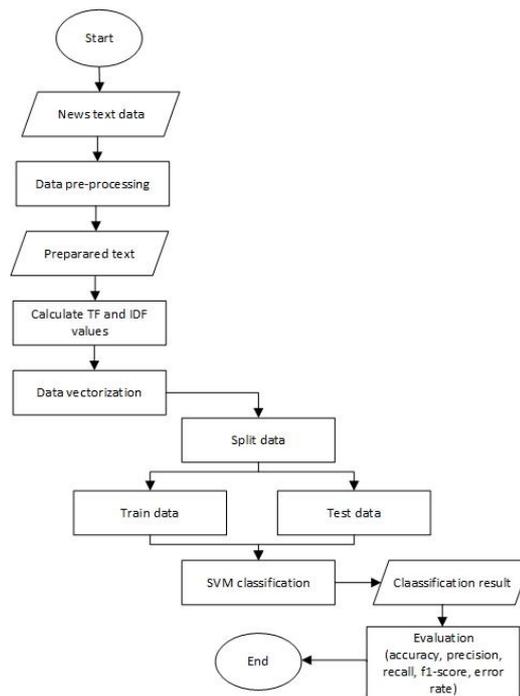
$$Recall = \frac{tp}{tp+tn} \quad (5)$$

F-measure can be calculated with the formula:

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (6)$$

The creation of a binary classification model is the first step in this research that can distinguish hoax news and real news with the SVM methods. The second stage is testing the model that has been built. As for the model requirements from observations on the literature that has been obtained, several needs are defined that must be met by a model in order to be said to be good, namely the model can produce accuracy above 75%, and can measure the value of recall, precision, F1 score, balanced error rate.

The classification process using SVM starts with data preprocessing. Then proceed with text vectorization. In this study, the vector comprises two parts: weights and dimensions (word id). This weight is combined into a TF-IDF value. Following vectorization, the stage is completed by dividing the data into training and test sets. The training data are used to build the SVM model, while the test data are used to assess the model's quality. The 10-fold cross validation approach evaluates the model's validity. Figure 1 below shows the model's operational steps.



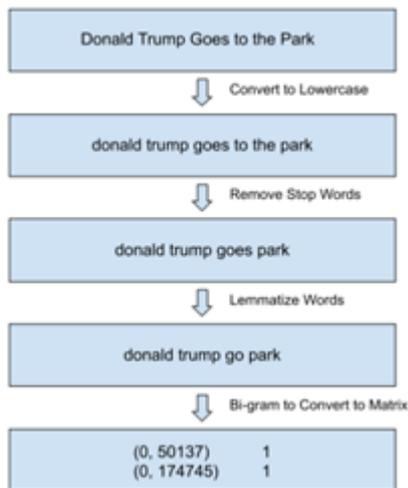
Picture 1. SVM Model Research Flow Chart (source: private)

Fake news data is downloaded from the link <https://www.kaggle.com/datasets/clmentbisailon/fake-and-real-news-dataset> in 2020 with a total of 44,907 lines of data divided into two pieces of data as much as 21,418 lines are real news and 23,489 lines are hoax news data and each data contains headlines, news topics, and news years. The second data comes from scrapping on the BBC website as much as 1574 pieces of data that have been labeled.

The following are the stages of text data preprocessing:

1) Case Folding

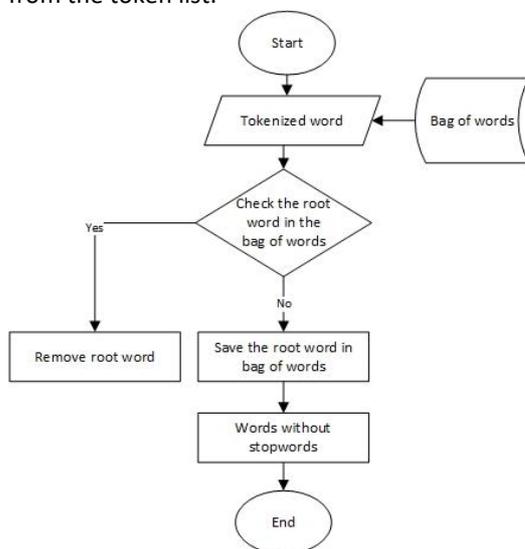
The initial step in document preprocessing is case folding. All of the document's letters are changed to lowercase during this process. In this process, only letters a through z are accepted. The procedure of folding a case is depicted in Picture 2 below.



Picture 2. Case Folding Process[9]

1) Stopword Removal

The stopword elimination stage is where the text's unnecessary words are eliminated. Stopwords are words that are not descriptive and can be eliminated using the bag-of-words technique. Picture 3 shows the flowchart of stopword removal. In the stopword removal module, each root word will be checked by the system whether it is in the stopword dictionary or not. If it is, the root word will be removed from the token list.



Picture 3. Stopword Removal Process (source: private)

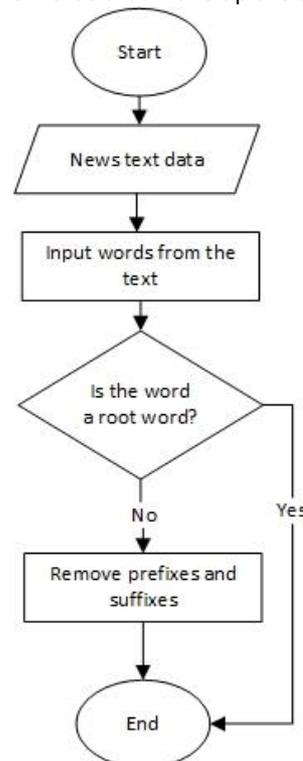
2) Stemming

The process of locating the root word for each word in a filtered result is known as stemming. At this point, diverse word formations are converted back into the same representation.

The basic word search is done by removing all affixes from the word, be it prefixes, inserts, or suffixes. Picture 4 shows the flowchart of stemming. In this design will use TF-IDF for Python language in the stemming process.

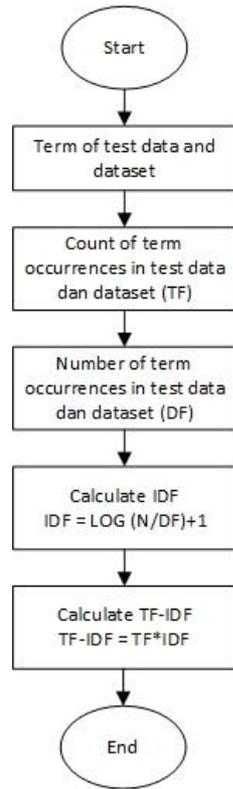
3) Tokenizing

Tokenizing is the process of dividing the input string into its individual words. Characters other than letters will be considered delimiters and will be removed or deleted for the process of getting the words that make up the text.



Picture 4. Stemming Process (source: private)

The process continues by calculating TF-IDF. In order to determine the probability of occurrence of each word in each document in the dataset, the training data and test data documents are weighted by words or terms at this point. The TF-IDF process uses the help of a model from `sklearn.feature_extraction.text`. Picture 5 shows the flowchart of the weight calculation process based on the TF-IDF value.

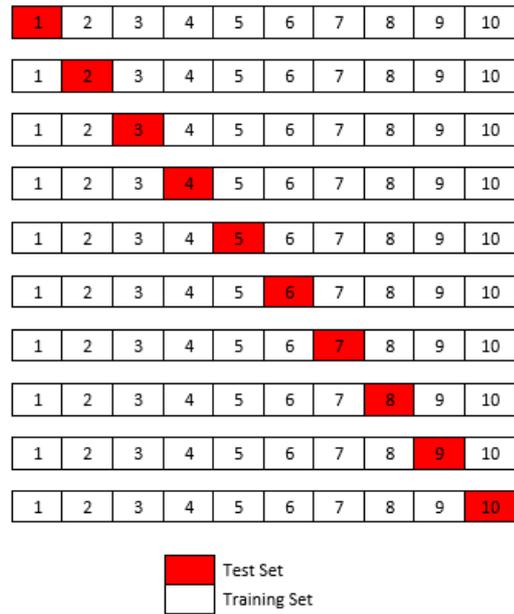


Picture 5. TF-IDF Calculation Process (source: private)

The vectorized data will be split into training data and test data during this SVM classification stage. The SVM model is built from the training data. Then predictions are made on test data.

The assessment stage is completed to assess the modeling's correctness after being applied to the training set of data. Then compare the results of two different datasets by applying Confusion Matrix to calculate precision, recall, f-measure score, and accuracy.

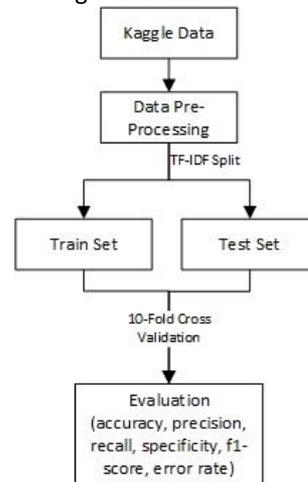
The K-Folds Cross Validation method, one approach for evaluating/validating the accuracy of a model created based on a specific dataset, will be used for test validation. Testing uses a value of K = 10 which indicates the number of datasets, where the dataset is divided into 10 parts, namely K1, K2, K3, K4, K5, K6, ..., K10 with testing iterations carried out 10 times where each dataset will alternately become test and training data. The test is data from Kaggle with the amount of test data used as much as 12,000 data. Each dataset (K) amounts to 1,200 data in English and the rest of the data will be used for training data. The 10-Folds Cross Validation test scenario that will be carried out can be seen in Picture 7 below.



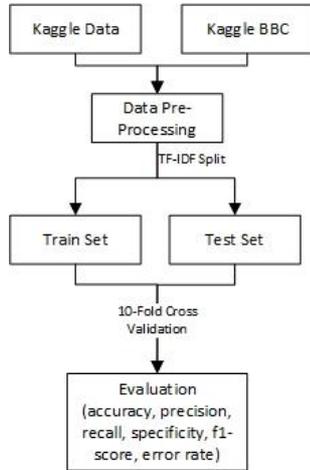
Picture 6. 10-Fold Cross Validation Testing Scheme [18]

In testing the test data used is data obtained from Kaggle and data obtained from the BBC site that has been scrapped with a total of 1500 pieces of data that have been labeled.

In this test scenario, model testing will be divided into two, namely for data obtained from Kaggle and data on BBC. In the first scenario, Kaggle data that has undergone data preprocessing and has been divided into training data and test data is tested with 10 fold cross validation so that a model's accuracy may be known. In the second scenario with BBC data, newly obtained data will be combined with training data from Kaggle data and then tested with the SVM method. The data testing scenario is depicted in Figures 8 and 9 below.



Picture 8. Test Scenario Using Kaggle Data (source: private)



Picture 9. Test Scenario Using BBC Data (source: private)

Fold	Accuracy
9	99,3%
10	98,8%

After the model goes through the training stage, the next step is to test the model with test data, the data used is 30% of 40,000 article data where the test results obtained by classification results and confusion matrix show that the model had an F1-score of 99% and had accuracy, recall, and precision values of 99.4%, 99.49%, and 99% respectively.

Table 3: Classification Results of Kaggle Test Data

Label	Precision	Recall	F1-score	Accuracy
REAL	99%	99%	99%	99,4%
FAKE	99%	99%	99%	

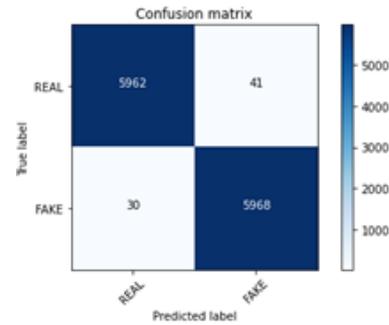
3. RESULT AND DISCUSSION

3.1 Model Testing Results of the first scenario with Kaggle Data

In the first test scenario with a dataset from Kaggle. 70 percent of the 40,000 article data, which included labeled original articles as well as fraudulent or hoax articles, were trained using 10-fold cross validation. From the training model that has been carried out, the model obtained the lowest accuracy result of 98.8% while the best accuracy obtained by the model was 99.6% contained in the 8th fold. Table 2 below shows the measurement results of the 10-fold cross validation.

Table 2: Model Training Results with 10 Cross-Fold Validation

Fold	Accuracy
1	99,5%
2	99,3%
3	99%
4	99,2%
5	99,2%
6	99,5%
7	99,3%
8	99,6%



Picture 9. Confusion Matrix of SVM Model Test Data (source: private)

Description of Confusion Matrix:

True Positive (TP) = 5962

True Negative (TN) = 5968

False Positive (FP) = 41

False Negative (FN) = 30

To get the Error-rate and Specificity values, it is calculated based on the value in the confusion matrix. The calculation of these values can be seen in the following calculation:

1) Error-rate:

$$\frac{fp+fn}{tp+fp+tn+fn} \times 100\%$$

$$= \frac{(41+30)}{(5962+41+5968+30)} \times 100\%$$

$$= \frac{71}{12001} \times 100\%$$

$$= 0,006 \times 100$$

$$= 0.6\%$$

2) Specificity:

$$\frac{tn}{(tn+fp)} \times 100\%$$

$$= \frac{5968}{(5968+41)} \times 100\%$$

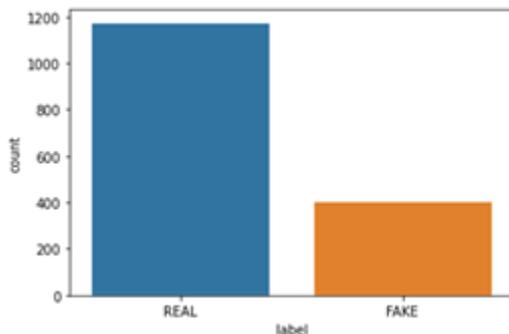
$$= \frac{5968}{6009} \times 100\%$$

$$= 0,9931 \times 100\%$$

$$= 99,31\%$$

3.2 Second Scenario Testing Results with BBC Data

In the second test scenario with a scraped dataset from the BBC in this test the amount of BBC data used was 1574 pieces of data consisting of 1172 real news data and 400 pieces of fake news data in English.



Picture 11. BBC Scrapping Data Count Chart (source: private)

According to the results of the testing done on the SVM model, the model has an accuracy of 98.9%, precision of 99%, recall of 99%, and F1-Score of 99%. The categorization findings and confusion matrix displayed in Table 4 and Picture 11 provide additional details.

Table 4: BBC Test Data Classification Results

Label	Precision	Recall	F1-score	Accuracy
REAL	99%	99%	99%	98.9%
FAKE	99%	99%	99%	

```
[ ] print(confusion_matrix(y_test, y_pred))
```

```
[[796 11]
 [ 5 761]]
```

Picture 11. Confusion Matrix of BBC Test Data

Confusion Matrix Description:

True Positive (TP) = 796

True Negative (TN) = 761

False Positive (FP) = 11

False Negative (FN) = 5

1) Error-rate:

$$\frac{fp+fn}{tp+fp+tn+fn} \times 100\%$$

$$= \frac{(11+5)}{(796+761+11+5)} \times 100\%$$

$$= \frac{16}{1573} \times 100\%$$

$$= 0.01 \times 100\%$$

$$= 1\%$$

2) Specificity:

$$\frac{tn}{(tn+fp)} \times 100\%$$

$$= \frac{761}{(761+11)} \times 100\%$$

$$= \frac{761}{772} \times 100\%$$

$$= 0.985 \times 100\%$$

$$= 98,5\%$$

Table 5: Model Training Results with 10 Cross-Fold Validation for BBC data

Fold	Accuracy
Fold 1	98.4%
Fold 2	98.2%
Fold 3	98.7%
Fold 4	98.5%
Fold 5	98.6%
Fold 6	98.8%
Fold 7	98.4%
Fold 8	98.1%
Fold 9	98.7%

Fold	Accuracy
Fold 10	98.5%

3.3 Model Testing Analysis

The results of testing and training that have been carried out by the model in classifying both news data show good accuracy in both test scenarios. This proves that the SVM method is able to detect English hoax news with an accuracy rate of 99%. Even with unbalanced data, the SVM method is still able to provide good accuracy in the second test scenario. The test results also show the ability to recognize balanced fake news and real news tuples, indicated by balanced recall and specificity values. The second test also proves that the SVM method provides high generalization, from Kaggle training data with BBC test data, this method is able to recognize fake news that is not included in the training data at all.

4. CONCLUSION

Based on the findings of this study, we can draw the conclusion that the use of the SVM algorithm for the classification of hoax/fake news in English is very successful. This is because the model developed successfully distinguishes between news that is classified as fake and real news with a high accuracy value of 99.4% and an error rate of 1%. The second conclusion is that the SVM approach has been shown to have great generalization, as seen by the model's high value of 98.9% in the classification process using data from the BBC.

STATEMENT OF APPRECIATION

Our appreciation goes to DRPM INSTIKI for its support for this research, both material and non-material support. Hopefully this research will be useful for further research.

REFERENCES

[1] M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, dan J. Vilares, "Sentiment Analysis for Fake News Detection," *Electron. 2021, Vol. 10, Page 1348*, vol. 10, no. 11, hal. 1348, Jun 2021, doi: 10.3390/ELECTRONICS10111348.

[2] Jasmi, "Mengenai Ciri-ciri Berita Hoax dan Dampaknya," *QuBisa*, 2021. <https://www.qubisa.com/article/ciri-ciri-dan-dampak-hoax>.

[3] J. A. Nasir, O. S. Khan, dan I. Varlamis, "Fake news detection: A hybrid CNN-RNN based deep learning approach," *Int. J. Inf. Manag. Data Insights*, vol. 1, no. 1, hal. 100007, Apr 2021, doi: 10.1016/J.JJIMEI.2020.100007.

[4] Fangyi Yu, "How to Build a Fake News Detection Web App Using Flask," *Towards Data Science*, 2021. <https://towardsdatascience.com/how-to-build-a-fake-news-detection-web-app-using-flask-c0cfd1d9c2d4> (diakses Des 08, 2022).

[5] J. L. Imbwaga, N. Chittaragi, dan S. Koolagudi, "Fake News Detection Using Machine Learning Algorithms," in *Proceedings of the 2022 Fourteenth International Conference on Contemporary Computing, 2022*, hal. 271–275, doi: 10.1145/3549206.3549256.

[6] I. Ahmad, M. Yousaf, S. Yousaf, dan M. O. Ahmad, "Fake News Detection Using Machine Learning Ensemble Methods," *Complexity*, vol. 2020, 2020, doi: 10.1155/2020/8885861.

[7] Z. Khanam, B. N. Alwasel, H. Sirafi, dan M. Rashid, "Fake News Detection Using Machine Learning Approaches," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1099, no. 1, hal. 012040, 2021, doi: 10.1088/1757-899x/1099/1/012040.

[8] P. S. Reddy, D. Elizabeth Roy, P. Manoj, M. Keerthana, dan P. V. Tijare, "A study on fake news detection using naïve bayes, SVM, neural networks and LSTM," *J. Adv. Res. Dyn. Control Syst.*, vol. 11, no. 6 Special Issue, hal. 942–947, 2019.

[9] B. Altman, R. Kim, J. Spsychalski, dan R. Spear, "DETECTING FAKE NEWS USING SUPPORT VECTOR MACHINES," 2021.

[10] S. Vinothkumar, S. Varadhaganapathy, M. Ramalingam, D. Ramkishore, S. Rithik, dan K. P. Tharanies, "Fake News Detection Using SVM Algorithm in Machine Learning," in *2022 International Conference on Computer Communication and Informatics (ICCCI)*, 2022, hal. 1–7, doi: 10.1109/ICCCI54379.2022.9740886.

[11] V. Gupta, R. S. Mathur, T. Bansal, dan A. Goyal, "Fake News Detection using Machine Learning," *2022 Int. Conf. Mach. Learn. Big Data, Cloud Parallel*

- Comput. COM-IT-CON 2022*, vol. 9, no. 6, hal. 84–89, 2022, doi: 10.1109/COM-IT-CON54601.2022.9850560.
- [12] S. Kumar, A. K. Kar, dan P. V. Ilavarasan, “Applications of text mining in services management: A systematic literature review,” *Int. J. Inf. Manag. Data Insights*, vol. 1, no. 1, hal. 100008, Apr 2021, doi: 10.1016/J.JJIMEI.2021.100008.
- [13] Algonz D.B. Raharja, “Machine Learning: Pengertian, Cara Kerja, dan 3 Metodenya,” *Ekrut Media*, 2022. <https://www.ekrut.com/media/apa-itu-machine-learning> (diakses Des 08, 2022).
- [14] K. N. Singh, S. D. Devi, H. M. Devi, dan A. K. Mahanta, “A novel approach for dimension reduction using word embedding: An enhanced text classification approach,” *Int. J. Inf. Manag. Data Insights*, vol. 2, no. 1, hal. 100061, Apr 2022, doi: 10.1016/J.JJIMEI.2022.100061.
- [15] M. Bansal, A. Goyal, dan A. Choudhary, “A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning,” *Decis. Anal. J.*, vol. 3, hal. 100071, Jun 2022, doi: 10.1016/J.DAJOUR.2022.100071.
- [16] Samsudiney, “Penjelasan Sederhana tentang Apa Itu SVM?,” *Medium*, 2019. <https://medium.com/@samsudiney/penjelasan-sederhana-tentang-apa-itu-svm-149fec72bd02>.
- [17] A. Fahrizain, “Bag of Words vs TFIDF — Penjelasan dan Perbedaannya |,” *Data Folks Indonesia*, Nov 01, 2021. <https://medium.com/data-folks-indonesia/bag-of-words-vs-tf-idf-penjelasan-dan-perbedaannya-3739f32cdc72> (diakses Mei 30, 2023).
- [18] Anonym, “10 Fold-Cross Validation,” *Binus University Graduated Program*, 2017. <https://mti.binus.ac.id/2017/11/24/10-fold-cross-validation/>.