Journal of Computer Science

Original Research Paper

Comparative Study: Algorithms for Short Message Service Classification

[1]Evaristus Didik Madyatmadja, [1]Aldi, [1]Fiona Fheren, [1]Helen Angelica, [1]Hanny Juwitasary and [2]David Jumpa Malem Sembiring

[1]Department of Information Systems, School of Information Systems, Bina Nusantara University, Jakarta, 11480, Indonesia
[2]Teknik Informatika, Institut Teknologi Dan Bisnis Indonesia, Medan, Indonesia

Article history
Received: 09-02-2023
Revised: 05-05-2023
Accepted: 25-07-2023

Corresponding Author:
Evaristus Didik Madyatmadja
Department of Information Systems, School of Information Systems, Bina Nusantara University, Jakarta, 11480, Indonesia
Email: emadyatmadja@binus.edu

Abstract: This research aims to classify Short Message Service (SMS) data by applying classification models that have studied SMS data to classify SMS data into SMS spam and SMS ham. The classification model is made from data mining algorithms: Naive Bayes and support vector machine. Before implementing the two algorithms, the SMS data will go through a text preprocessing stage, including data cleaning (whitespace removal, removal of punctuation, and removal of numbers), case folding, stemming, tokenizing, and stop word removal. In this research, a comparison of the accuracy of the two data mining methods will be carried out to see and get the best classification algorithm. Researchers also implemented several experiments by comparing the use of testing data by 20 and 30% and comparing the application of preprocessing stemming and without stemming. This study found that the support vector machine algorithm using testing data of 20% by applying the stemming stage had the highest accuracy rate, 97.5%.

Keywords: SMS Spam, SMS HAM, Naive Bayes, Support Vector Machine, Classification, Data Mining, Text Mining

Introduction

The COVID-19 pandemic began in 2020, encouraging Indonesians to adapt and move quickly in the face of changes, one related to communication and information technology. With work-from-home schemes and learning activities conducted online (learn from home), people are indirectly required to rely on technology, including telecommunications, to do their activities. According to data from the ministry of communication and Informatics, the number of smartphone users in Indonesia reached 167 million people (89% of the total population in Indonesia in early 2021) based on the investor. id page about the use of smartphones in Indonesia has increased by approximately 75 million users. The number of smartphone users increases the development of technology that occurs. The development of this technology can impact the growth of telecommunications to convey information remotely and with various parties wherever they are.

Problem Identification

However, many technological developments should be used more by responsible parties. One of the abuses of this technological development can be seen in the misuse of communication and information technology by using media such as SMS, email, telephone, and other sites and blogs containing offers or advertisements that direct users or targets to access unofficial untrusted sites. One of the problems that still needs to be more widely experienced by smartphone users is the use of SMS. Although many applications have replaced the role of SMS features in smartphones, it cannot be denied that many fraud modes are still found using SMS. Some types of fraud are categorized as SMS spam, such as SMS rewards from a certain agency, SMS Online Loan Offers, SMS OTP Code Request, SMS Marketing, to SMS Threats. SMS spam also occurs in users who just bought a phone number from an operator service and did not only occur in old users.

According to the True caller insights report 2019, Indonesia became the 10th country in the world to receive the most spam SMS, with the average number of spam SMS received by every mobile phone and smartphone user in Indonesia as many as 46 SPAM SMS every month. This is supported by a statement by the chairman of the Badan Perlindungan Konsumen (BPKN) assessment and development commission Arief safari. In contrast, as of August 2020, there have been 3,269 complaints related to SMS spam since 2017. The COVID-19 situation is one of

the reasons for those who feel interested in the proposal given because of the difficulties or economic obstacles they experience.

Research Questions

The following are some of the studies in this study: (1) Can the Naïve Bayes classifier algorithm and support vector machine be used to classify SMS spam? (2) What better algorithm is between the Naïve Bayes classifier and support vector machine for classifying SMS spam?

Research Benefits

Some ways that can distinguish whether SMS received is SMS spam or not. One of them is by utilizing data mining techniques to label SMS. This research was conducted with several predetermined restrictions: (1) SMS data received by some mobile phone users/smartphones, (2) SMS collected from 2019-2021. The following are some of the studies in this study: (1) Can the Naïve Bayes classifier algorithm, Support vector Machine, and decision tree be used to classify SMS spam? (2) What better algorithm is between the Naïve Bayes classifier, support vector machine, and decision tree for classifying SMS spam?

Data Mining

Data Mining is processing data from un pattern data into useful information or knowledge (Suntoro, 2019). Data mining has significance in finding patterns, forecasting and discovering knowledge, etc., in various business fields (Pandiangan *et al*., 2020). Three learning methods are used in data mining: Supervised, unsupervised, and semi-supervised. In the managed learning method, it is necessary to determine the input variable used as the correct label to be learned to make it easier for the algorithm to create a model for performing classification and regression. For the unsupervised learning method, the target variable is not determined to look for an unknown pattern, which is usually used in clustering and association. As for the semi-supervised learning method, the target variable is only partially and does not have large data. The forms that can be used include classification, regression, and prediction. There is no best method, or one size fits all. Finding the correct algorithm is partly just trial and error; even highly experienced data scientists can't tell whether an algorithm will work without trying it out (Kavitha and Elango, 2017).

There are several methods used to analyze data to produce valuable information, including (1) Classification, the method used in data mining, where this method groups criteria variables with target variables to predict labels, (2) Clustering, methods used to group similar values, (3) Association, a process of finding attributes that have the same frequency of occurrence, (4) Regression, a method that resembles classification but does not use a predetermined label to be used in finding patterns, but by using variables dependent as criteria and independent variables as predictors by assuming the model.

Text Mining

Text mining is used to extract useful information, usually in documents, by identifying and looking for patterns. In text mining, the information obtained is unstructured, so it requires converting text data into structured data and simplifying the text preprocessing analysis process. Several steps are used to perform text preprocessing: Case folding, tokenizing, stemming, and stop word removal.

Classification

Classification is a method used in data mining to analyze and produce valuable information. The classification works by grouping the data based on the target and criterion variables to predict the label. This method has two stages, learning and classification; during the learning process, it will conduct data training using an algorithm which will then be continued in the classification process, where the algorithm that has been trained will test the data to ensure the accuracy of the pattern as a result of the learning.

Naive Bayes Classifier

Naïve Bayes is an algorithm that is often used for classification. This Naïve Bayes algorithm is predictive and descriptive, analyzing the relationship between the dependent and independent variables to generate the probability of each possibility. This algorithm requires two datasets, a training dataset and a testing dataset. The Naïve Bayes algorithm is often used because the amount of data needed is minimal and only requires a small amount of training data. It can also be used for qualitative and quantitative data, easier to understand and implement than other algorithms.

The equation of Naïve Bayes is as follows:

$$P(Ci\,|X) = \frac{P(X\,|Ci)p(Ci))}{P(x)} \quad (1)$$

X	=	Criteria for a case based on input
Ci	=	The solution class of the i^{th} pattern, where i is the number of class labels
$P(Ci \mid X)$	=	Probability of class Ci label with input criteria X
$P(X \mid Ci)$	=	Probability of input criteria X with class label Ci
$P(Ci)$	=	Probability of class label Ci

Support Vector Machine

Support Vector Machine (SVM) is an algorithm used in data mining to classify by finding the best hyperplane. The hyperplane is a function that distinguishes between two classes to classify the tested data. This algorithm can solve both linear and non-linear problems by using kernel functions. Initially, SVM worked only on linear problems and was developed for non-linear problems using kernel functions. This kernel function takes a low-dimensional input space or an issue that was initially inseparable and then converts it to a higher-dimensional one or separates the problem by adding more dimensions. The effectiveness of SVM is often influenced by the type of kernel function selected and tuned based on the characteristics of the data (Haddi *et al*., 2013).

The SVM algorithm can be implemented in the Python programming language with the SVM. SVC library has several kernel parameters. For example, there are linear kernels, polynomial kernels, and RBF kernels. From the research conducted by Neli Kalcheva, Milena Karova, and Ivaylo Penev, several SVM classification experiments with Kernel Parameters were carried out, namely linear kernel, poly kernel, RBF kernel, and sigmoid kernel (Kalcheva *et al*., 2020). This study found that the Linear Kernel has the highest accuracy value in the shortest time.

In a study by Luo (2021) on English Text Classification, the SVM algorithm has higher efficiency and accuracy than Naive Bayes and logistic regression, which is based on several parameters, namely precision, recall, and F1-score.

The advantage of this SVM algorithm is the result of its higher accuracy value than other algorithms, runs better on higher dimensions, and is easier to implement because the determination of the support vector can be applied to the Quadratic Programming Problem (QP Problem). The disadvantage is that it is difficult to process when you have a huge amount of data and in theory, it can only process two classes. This disadvantage is evidenced by research conducted by Miao *et al*. (2018). The results showed that SVM requires more processing time than naive Bayes and K-nearest neighbor algorithms. SVM takes about 484.75s, Naive Bayes takes about 3.42 s and K-Nearest neighbor takes about 19.60 s.

Decision Tree

The decision tree is a classification method that converts facts into a decision tree form that represents a rule or pattern. This decision tree also serves to help explore data and find hidden patterns or relationships between attributes or data. A decision tree in the context of classification can be called a classification tree. Many algorithms can be used in forming a decision tree, including C4.5, ID3, and CART.

In the decision tree, there is a node that represents the root and leaf. The source will be connected to the leaf, which is represented in a line that is analogous to a branch of the tree itself. The root node is located at the top of the decision tree and can only have one root node. An internal node is a branch that starts from the root node. The leaf node is the decision tree's final node, which has only one input and no output.

The decision tree has several advantages, such as, compared to other algorithms, data preparation during preprocessing can be done without much effort. It does not need data normalization and the way these algorithms work is very easy to understand, so it is easier to interpret. Meanwhile, the disadvantage of using this algorithm is that when there is a small data change, it will significantly affect the results that will be issued.

SMS

SMS or Short Message Service is one of the technology services used to exchange messages via mobile devices such as mobile phones (Tubagus, 2018). The convenience when using SMS is that we don't have to bother to buy internet data so we can directly use mobile credit. The cost of sending SMS for each operator in Indonesia also varies depending on the use of the operator and it can also be accessible to other operators. The way SMS works is using the store and forward method (Ardiansyah, 2017), meaning that the sender and recipient do not have to be on the network when sending and receiving messages because the message is forwarded by the sender to the Short Message Service Center (SMSC) which will then receive and forward the message back (Ardiansyah, 2017). When the SMS recipient is ready and on the network at a later time, there is no need for direct transmission from sender to recipient.

This study will use two types of SMS text: SMS spam and spam. Spam is sending messages to other people without being desired by the recipient by using electronic devices continuously and usually in large amounts and very disturbing; the message sent is called SMS spam. This is because it occupies storage space in the phone and computational power (Sravya and Pradeepini, 2020).

While ham is an SMS that is not spam or also called non-spam so that the recipient wants to receive the message, Ham SMS usually contains statements that are not malicious. In general, ham texting tends to lead to important messages to people you know and have texted each other in the past.

Materials

In this pandemic, many irresponsible parties are taking advantage of the situation by attracting the attention of mobile network users or potential targets by sending spam SMS related to online loans and/or lotteries. Although

many parties have often ignored SMS spam, unfortunately, many people still trust the offers provided by irresponsible parties through SMS spam. The COVID-19 situation is one of the reasons for those who feel interested in the proposal given because of the difficulties or economic obstacles they experience.

Methods

In the process of researching to detect spam messages, there are several stages that will be passed by researchers, namely in the following image.

Research Stages

The picture on Fig. 1 shows the process of processing SMS data to be studied using the Naive Bayes classifier, Support Vector Machine (SVM), and decision tree algorithms. It aims to classify SMS either into spam or ham categories.

Method of Collecting Data

In general, when conducting research, researchers will undoubtedly look for data sources that will be used to detect SMS spam by looking at the results of the accuracy values of the Naive Bayes classifier, Support Vector Machine (SVM), and decision tree algorithms. The data source that will be used is SMS with spam and spam categories. The total SMS data used to classify SMS is about 1000, which includes SMS with spam and ham categories. Before entering the stage of using data mining algorithms, namely Naive Bayes classifier, Support Vector Machine (SVM), and decision tree, researchers will go through the text preprocessing stage to change SMS data to a more structured format.

Text Preprocessing

Text mining is the initial stage where the process is used to make text data more structured and feasible to be processed at a later stage. This text preprocessing stage uses Google Colab with the Python programming language. In this study, there were six stages carried out, namely:

a. Data cleaning
   - Whitespace removal
     Whitespace removal is used in dirty data as a process to remove excess spaces at the beginning and end of sentences in text data. For example, "thesis" becomes "thesis"
   - Punctuation removal
     Punctuation removal is the process of removing punctuation marks, such as. ,! “ $ # & * and so on
   - Remove number
     Remove number is a process to remove the digits contained in text data

b. Case folding: Case folding is converting all letters in a word into lowercase form
c. Stemming: Stemming is the process of returning a word to its basic form. The changed words are words that contain affixes
d. Tokenizing: Tokenizing is converting a sentence into a token by separating or splitting sentences into words. For example, “I received spam SMS” becomes ['I', 'received,' 'SMS', 'spam']
e. Stopword Removal: Stopword removal is a process to remove all words that are considered unimportant text data. Dirty data is defined as inaccurate, inconsistent, and incomplete due to the error found within the dataset (Ridzuan and Zainon, 2019). In this research, the stop words removal stage was carried out using the Python programming language and Sastrawi library, namely, stop words words (‘Indonesian’)

Data Splitting

In the data splitting stage, the researcher divides the SMS data randomly into two parts, where some SMS data is training data and the rest is testing data. The definition of training data is data that is used as learning to get the suitable class or label while testing data will be used as a test to see whether the precision and accuracy of the results of class learning or labels are in accordance with the data that has been previously trained.

When dividing SMS data into several parts, a certain method will be used so that the SMS data is divided according to the size or ratio that has been determined. The method used for data splitting is Simple Random Sampling. This method is the most used. Simple random sampling is an essential technique in sampling where each element (n) taken from a population (N) has an equal chance. It is stated as simple (simple) because the sampling of population members is done randomly without regard to the strata that exist in the population (Golzar *et al*., 2022). Therefore, simple Random Sampling is considered one of the fairest methods of selecting samples from the population (Sharma, 2017).

TF-IDF

TF-IDF is one of the feature extraction methods used to generate new data sets but with a smaller amount of data than the original data. The TF/IDF method was chosen to be a vectorization representation of text data because it works nicely to improve the performance value of the classification model, which is related to recall and precision values so; that in the end, this method is believed to provide more accurate results (Prasetijo *et al*., 2017). Hoax detection system on Indonesian news sites based on text classification using SVM and SGD) (Prasetijo *et al*., 2017).

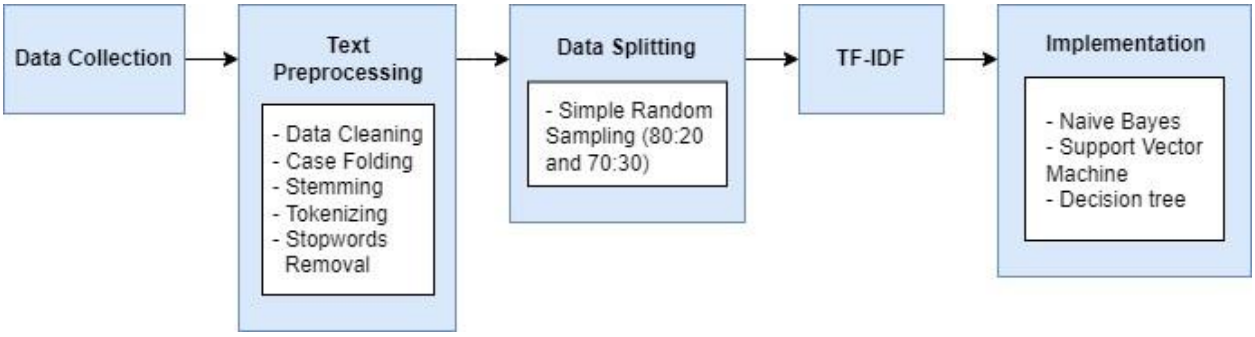


Fig. 1: Research stages

From the previous research conducted by Melvin Diale, Christiaan van der Walt, Turgay Celik, and Abiodun Modupe related to email spam filtering, it was mentioned that the Support Vector Machine classification algorithm performed better with the application of feature extraction on the email body. In contrast, the Naive Bayes classification algorithm had a weaker performance when applying feature extraction to the email body (Diale *et al*., 2016).

At this stage, the researcher uses the TF-IDF algorithm to calculate the value of each word in a sentence or text. The TF/IDF method was chosen to be a vectorization representation of text data because it works nicely to improve the performance value of the classification model related to recall and precision values. In the end, this method is believed to provide more accurate results. From this stage, it is possible to categorize whether an SMS text is included in the SMS spam or ham category based on the keywords of the messages received by the user that show on Fig. 2. Again, this method is believed to provide accurate results. Figure 3 show the calculation on TF/IDF is the value of Term Frequency (TF) and Inverse Document Frequency (IDF) for each token or word in the corpus or a collection of texts or documents. Here is the formula for calculating the TF-IDF:

$$TF * IDF = TF * Log\frac{n}{df}$$

TF = Term frequency
df = Document frequency
n = Number of the document

The importance of a word in a document depends on the frequency with which the word appears in the document. The frequency of documents in which more and more words appear indicates how common the word is.

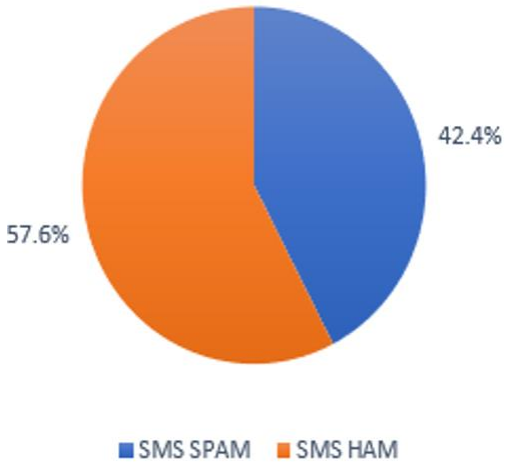


Fig. 2: SMS data category percentage implementation results

```
print(train_y_tfidf)
print(test_y_tfidf)
```

```
  (799, 903)    0.27467863258507114
  (799, 833)    0.20790614314587677
  (799, 434)    0.32089554134154447
  (799, 420)    0.26416405513031305
  (799, 192)    0.25557301981115778
  (799, 46)     0.27467863258507114
  (799, 19)     0.28097066371087875
  (0, 2588)     0.29873957764605
  (0, 2495)     0.12054327601696395
  (0, 2225)     0.19416971320672833
  (0, 1968)     0.13320663082022205
  (0, 1854)     0.18662126158613412
  (0, 1656)     0.16078011208812182
  (0, 1643)     0.29873957764605
  (0, 962)      0.09721459641542683
  (0, 620)      0.17269068505711355
  (0, 451)      0.14206708714582988
  (0, 395)      0.20777039268923855
  (0, 357)      0.2281598353169265
  (0, 157)      0.5598637847584023
  (0, 62)       0.29873957764605
  (0, 59)       0.36535048567704054
  (1, 2425)     0.5070328028092785
  (1, 2411)     0.32158914940994133
  (1, 1851)     0.23144005389357938
  (1, 1820)     0.27808994569199133
  (1, 1742)     0.40961942412740955
```

Fig. 3: TF-IDF first experiment (80:20) with stemming

Results

TF-IDF

DOI: 10.3844/jcssp.2023.1333.1344

Confusion Matrix (Simple Random Sampling)

1. Naïve Bayes

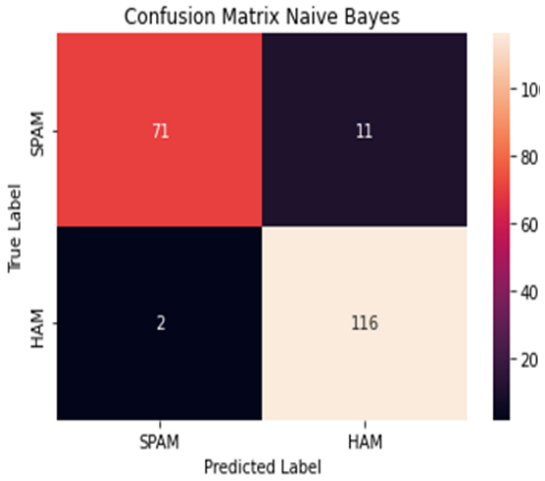


Fig. 4: Naïve Bayes first experiment (80:20) with stemming

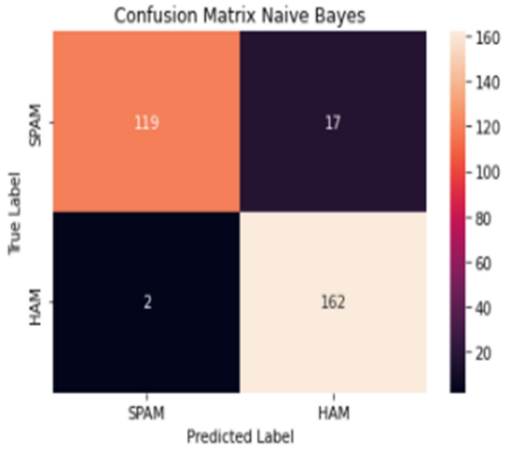


Fig. 5: Naïve Bayes second experiment (70:30) with stemming

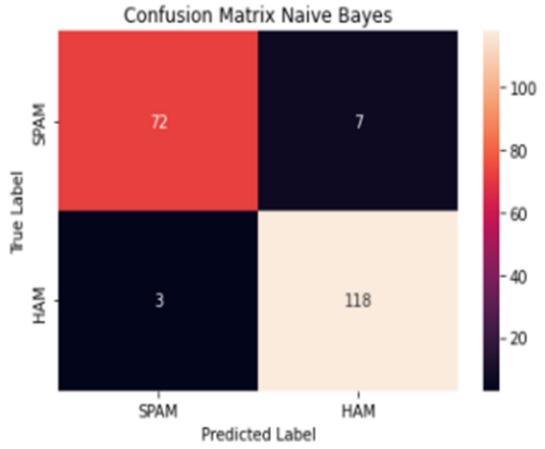


Fig. 6: Naïve Bayes third experiment (80:20) without stemming

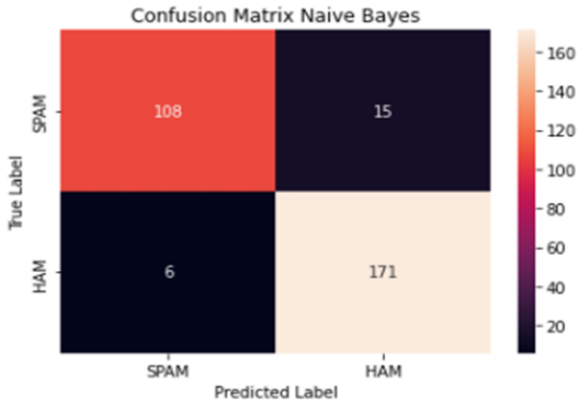
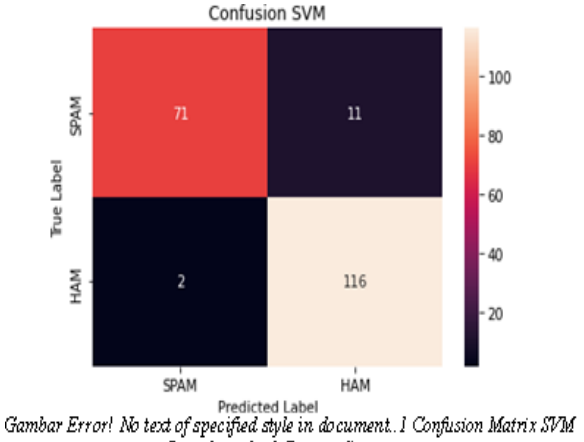


Fig. 7: Naïve Bayes fourth experiment (70:30) without stemming

Grouping using the Naive Bayes classifier as depicted in Figs. 4-7 is used to determine which class is the most optimal.

2. Support Vector Machine



Gambar Error! No text of specified style in document..1 Confusion Matrix SVM Percobaan ke-1 Dengan Stemming

Fig. 8: Support vector machine first experiment (80:20) with stemming

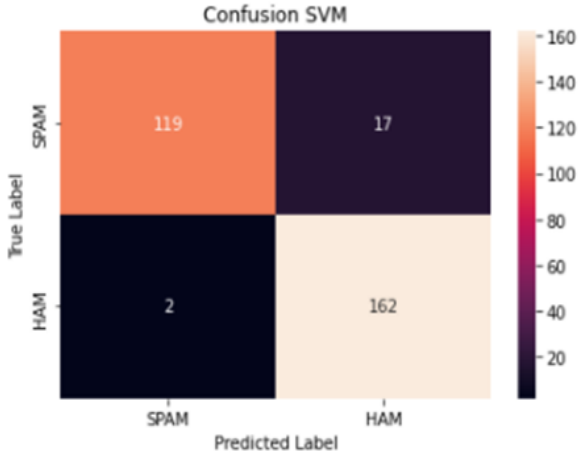


Fig. 9: Support vector machine second experiment (70:30) with stemming

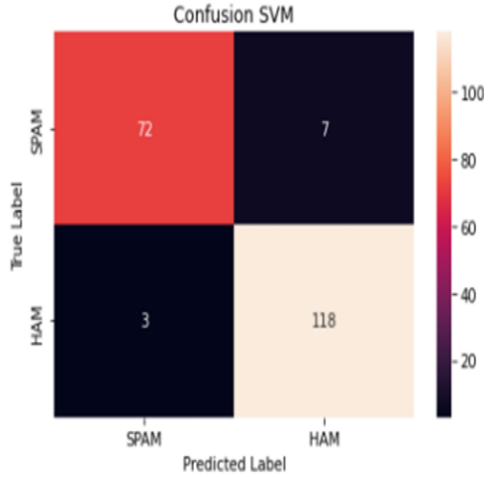


Fig. 10: Support vector machine third experiment (80:20) without stemming

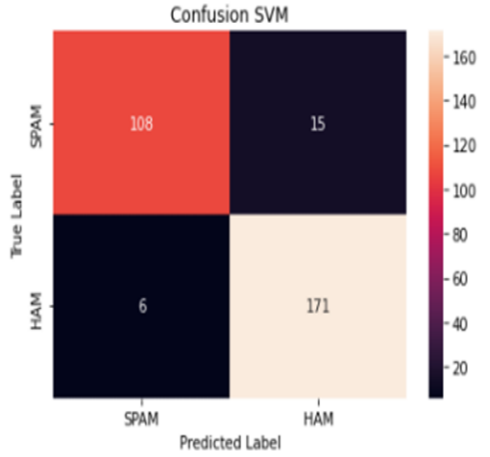


Fig. 11: Support vector machine fourth experiment (70:30) without stemming

Support vector machine experience to get a higher accuracy value that show on Figs. 8-11.

3. Decision Three First

The result of experiment using decision tree first shown on Figs. 12-15.

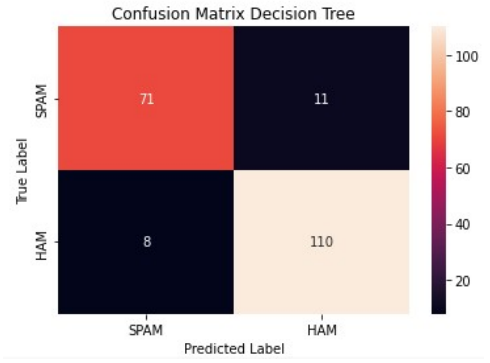


Fig. 12: Decision tree first experiment (80:20) with stemming

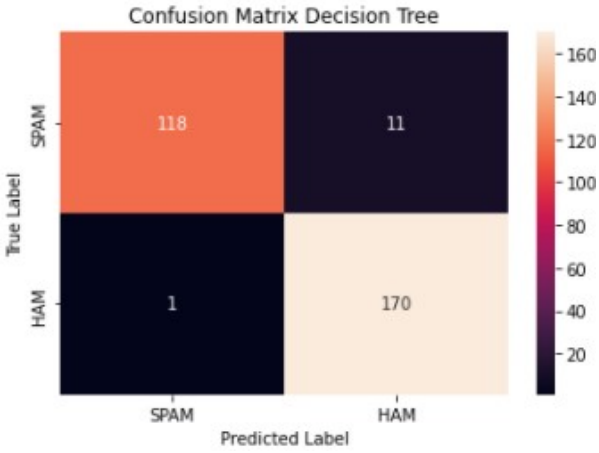


Fig. 13: Decision tree second experiment (70:30) with stemming

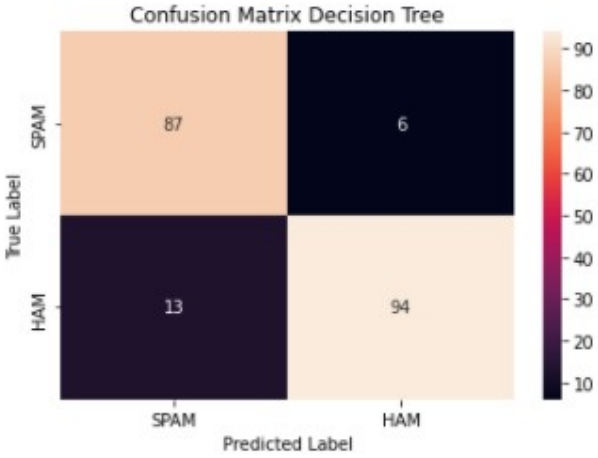


Fig. 14: Decision tree third experiment (80:20) without stemming

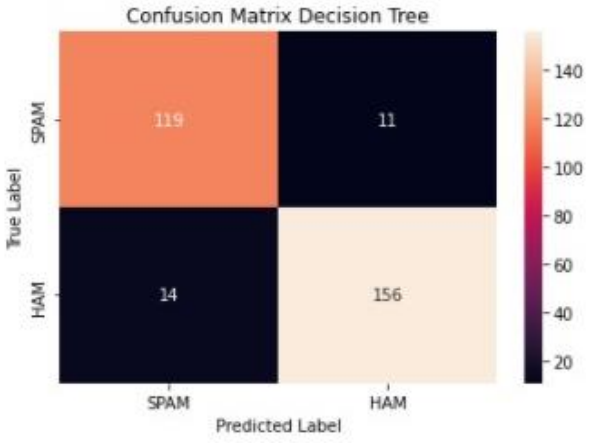


Fig. 15: Decision tree fourth experiment (70:30) without stemming

Accuracy Per Algorithm

1. Naïve Bayes

Table 1: Detail accuracy Naïve Bayes first experiment (80:20) with stemming

	Naive Bayes's first experiment			
	Precision %	Recall %	F1-score %	Support
HAM	97	87	92	82
SPAM	91	98	95	118
Average				
Macro average	94	92	93	200
Weighted average	94	94	93	200

Table 2: Detail accuracy Naïve Bayes second experiment (70:30) with stemming

	Naive Bayes's second experiment			
	Precision %	Recall %	F1-score %	Support
HAM	98	88	93	136
SPAM	91	99	94	164
Average				
Macro average	94	93	94	300
Weighted average	94	94	94	300

Table 3: Detail accuracy Naïve Bayes third experiment (80:20) without stemming Naive Bayes third experiment

	Precision %	Recall %	F1-score %	Support
HAM	96	91	94	79
SPAM	94	98	96	121
Average				
Macro average	95	94	95	200
Weighted average	95	95	95	200

Table 4: Detail accuracy Naïve Bayes fourth experiment (70:30) without stemming

	Naive Bayes's fourth experiment			
	Precision %	Recall %	F1-score %	Support
HAM	95	88	91	123
SPAM	92	97	94	117
Average				
Macro average	93	92	93	300
Weighted average	93	93	93	300

2. Support Vector Machine

Table 5: Detail accuracy support vector machine first experiment (80:20) with stemming

	Support vector machine first experiment			
	Precision %	Recall %	F1-score %	Support
HAM	96	98	97	82
SPAM	98	97	98	118
Average				
Macro average	97	98	97	200
Weighted average	98	97	98	200

Table 6: Detail accuracy support vector machine second experiment (70:30) with stemming

	Support vector machine second experiment			
	Precision %	Recall %	F1-score %	Support
HAM	96	97	97	136
SPAM	98	97	97	164
Average				
Macro average	97	97	97	300
Weighted average	97	97	97	300

Table 7: Detail accuracy support vector machine third experiment (80:20) without stemming

	Support vector machine third experiment			
	Precision %	Recall %	F1-score %	Support
HAM	92	97	94	79
SPAM	98	94	96	121
Average				
Macro average	95	96	95	200
Weighted average	96	95	96	200

Table 8: Detail accuracy support vector machine fourth experiment (70:30) without stemming

	Support vector machine fourth experiment			
	Precision %	Recall %	F1-score %	Support
HAM	93	95	94	123
SPAM	97	95	96	117
Average				
Macro average	95	95	95	300
Weighted average	95	95	95	300

3. Decision Tree

Table 9: Detail accuracy decision tree first experiment (80:20) with stemming

	Decision tree first experiment			
	Precision %	Recall %	F1-score %	Support
HAM	90	87	88	82
SPAM	91	93	92	118
Average				
Macro average	90	90	90	200
Weighted average	90	91	90	200

Table 10: Detail accuracy decision tree second experiment (70:30) with stemming

	Decision tree second experiment			
	Precision %	Recall %	F1-score %	Support
HAM	91	96	94	129
SPAM	97	93	95	171
Average				
Macro average	94	95	94	300
Weighted average	94	94	94	300

Table 11: Detail accuracy decision tree third experiment (80:20) without stemming

	Decision tree third experiment			
	Precision %	Recall %	F1-score %	Support
HAM	87	94	90	93
SPAM	94	88	91%	107
Average				
Macro average	91	91	90	200
Weighted average	91	91	91	200

Table 12: Detail accuracy decision tree fourth experiment (70:30) without stemming

	Decision tree fourth experiment			
	Precision %	Recall %	F1-score %	Support
HAM	89	92	90	103
SPAM	93	92	93	170
Average				
Macro average	91	92	92	300
Weighted average	92	92	92	300

Comparison Table of Accuracy Values

1. Naïve Bayes

Table 13: Accuracy comparison Naïve Bayes

Experiment	Accuracy (%)
First experiment (80:20) with stemming	93.50
Second experiment (70:30) with stemming	93.67
Third experiment (80:20) without stemming	95.00
Fourth experiment (70:30) without stemming	93.00

2. Support Vector Machine

Table 14: Accuracy comparison support vector machine

Experiment	Accuracy (%)
First experiment (80:20) with stemming	97.5
Second experiment (70:30) with stemming	97.0
The third experiment (80:20) without stemming	95.5
Fourth experiment (70:30) without stemming	95.0

Table 15: Accuracy comparison decision tree

Experiment	Accuracy (%)
First experiment (80:20) with stemming	90.50
Second experiment (70:30) with stemming	94.34
The third experiment (80:20) without stemming	90.50
Fourth experiment (70:30) without stemming	91.67

Some experiments using the Naive Bayes algorithm obtained the above results. The Table 1 shows the first experiment, Table 2 shows the second experiment, and Table 3 shows the experiment that has the highest accuracy value, namely in the third experiment with a comparison of training data and data testing which is 80:20 without applying stemming which has an accuracy value of 95%. And when do the fourth experiment on Table 4 shows the comparison of training data and data testing is 70:30.

After experimenting using the Support Vector Machine algorithm, there are results obtained in the table. The Tables 5-8 shows that the experiment with the highest accuracy score is the first experiment (80:20) with a stemming of 97.5%.

In several experiments using the Decision Tree algorithm, the results are shown in the Tables 9-12. The table shows the experiment that has the highest accuracy value, namely in the second experiment with a comparison of training data and testing data, which is 70:30, by applying stemming, which has an accuracy value of 94.34%.

Experiments using the Naïve Bayes Classifier algorithm (Figs. 4-7), support vector machine, and decision tree show that the accuracy value using the Support Vector Machine algorithm has a higher accuracy value. Of all the experiments resulting from using the support vector machine algorithm that shows on Tables 13-15, the first experiment (80:20) with Stemming resulted in the highest accuracy value of 97.5%. This shows that SMS classification using the support vector machine algorithm by applying text preprocessing stemming stages and dividing data into 80% data training and 20% data testing results in better SMS spam and ham classification compared to using stemming preprocessing text stages and dividing the data into 70% training data and 30% testing data.

The accuracy generated using the decision tree, Naive Bayes, and K-means algorithm models in Table 3 results in a high enough accuracy value proving that the data used is suitable.

Discussion

In several research studies, the text used in the classification process must be preprocessed first, consisting of data cleansing, case folding, stemming, tokenizing, and stop word removal. The stages of text preprocessing with several dimensions were carried out by researchers to be able to produce a better pattern/classification model to increase the level of accuracy of the algorithm used. From several previous studies, the data splitting stage with different ratios of training data and testing data will produce different levels of accuracy for each algorithm used.

The researcher conducted two data-splitting experiments in the research that the researcher did. The first experiment was carried out with a ratio of training data and testing data of 80:20 and the second experiment with a ratio of 70:30. The difference in accuracy results from different data splitting ratios is evidenced by research conducted by Madyatmadja *et al*. (2022) and research conducted by Zhao (2020). Furthermore, the research that the researcher did proves that using testing data as much as 20% will give a higher accuracy value than using data testing as much as 30%.

Another research that researchers found regarding the use of stemming in text preprocessing was in a journal entitled Improving the Accuracy of Text Classification using the stemming method, a case of nonformal Indonesian conversations where this journal tested the results of stemming between "Sastrawi" and "Incorbiz" using support vector machine algorithm on informal text data in Indonesian. The results of stemming accuracy with "Incorbiz" have a higher accuracy value of 85% compared to stemming with "Sastrawi" only 73%. Compared to the research that the researcher did using the SMS dataset, the study's results found that the highest accuracy value was in the use of no stemming in text preprocessing with a value of 95% on the Naive Bayes algorithm and support vector machine (Rianto *et al*., 2021). Improving the accuracy of text classification using the steaming method, a case of non-formal Indonesian conversation

Other research that supports the researcher's use of SMS data as an object of study is found in a journal entitled comparison performance of Naive Bayes Classifier and support vector machine algorithm for Twitter's classification of Tokopedia services (Kusumawati *et al*., 2019), where this research uses tweet data from twitter to see public opinion regarding services on Tokopedia. This aims to determine whether the opinion is categorized as positive sentiment or negative sentiment by using the Naive Bayes classifier and support vector machine algorithms. From the results of research using tweet data, the highest accuracy value is in the support vector machine algorithm at 83.34%. In comparison, the Naive Bayes classifier algorithm only gets an accuracy value of 75%. Then in the research that the researcher did use SMS data, the highest accuracy value was also generated by the support vector machine algorithm with a value of 97.5%. At the same time, the Naive Bayes classifier algorithm only had an accuracy value of 95%.

The research conducted by Ardianto (2020) regarding the measurement of opinion or separating positive and negative sentiments towards e-sports education on social media, namely Twitter. It was found that the Naive Bayes algorithm optimized using the Smote method has a higher accuracy value with a value of 70.32% compared to the support vector machine algorithm optimized using the Smote method, which has an accuracy value of 66.92%.

Researchers also found a journal titled Implementation of decision tree and Naïve Bayes classification method for Predicting study period which discusses the results of the comparison of using decision tree and Naive Bayes classification algorithms to improve student performance, one of which is student learning period. It is found that the decision tree algorithm has the highest accuracy value of 90% with a data splitting ratio of 60:40. The comparison of the use of training data does not affect the accuracy of the two algorithms used.

Implementation of the Dataset

After making a model for SMS data classification by conducting several experiments, the next step is for the researchers to implement a model for SMS data that does not have a variety or label. The model used in this stage is the classification model generated from the first experiment with the support vector machine algorithm using the simple random sampling data splitting method with a ratio of 80:20 which is the best experiment. In implementing the model on SMS that has yet to be classified or labeled, it will also go through the stages of text preprocessing and TF-IDF calculations. From the application of the first experiment model with the support vector machine algorithm on 420 SMS data that has not been classified or not yet labeled, the results obtained are 242, which are categorized as ham SMS and 178 as spam SMS.

Conclusion

Result

Based on this research, the object being used is the Indonesian language SMS data which contains messages about marketing, spam, ham, and others related to SMS. For SMS that refers to SMS spam, the researcher classifies it as a spam category. In contrast, for SMS that refers to other than SMS spam, the researcher will categorize it as ham category or commonly referred to as non-spam. SMS took from 2020-2021 with a total of 1000 SMS data.

The SMS data taken has passed the text preprocessing stage, which includes data cleaning, case folding, stemming, tokenizing, and stop word removal. After going through the text preprocessing stage, it will proceed to the data splitting to obtain training data and testing data used in determining the accuracy value results. The next stage is the term weighting stage, which is the TF-IDF calculation to get the weighted results for each term.

The best combination of these experiments that produces the highest accuracy level is using the support vector machine algorithm on a data splitting comparison of 80:20 and through the steaming process. This experiment got the highest accuracy rate of 97.5% with a precision value of 97%, recall of 97%, and an F1- score of 97%.

Recommendation

The suggestions from the thesis research that have been carried out by researchers to be investigated further are as follows:

1. Build machine learning to make it easier to classify SMS
2. Implement datasets other than SMS, such as email or WhatsApp

3. Using algorithms other than Naive Bayes classifier, support vector machine, and decision tree for better accuracy results
4. Implementing algorithms on datasets using other languages, such as English, whereby using that language, researchers can try using text preprocessing with the lemmatization stage
5. Added the amount of training data and test data to get better accuracy results
6. Changing Indonesian slang or slang words into standard language can help improve higher accuracy results
7. Using the Indonesian language corpus to convert non-standard language into the standard language

Acknowledgment

This study is supported by Research and Technology Transfer Office, Bina Nusantara University. The authors would like to thank all the researchers who provided the essential references and data for this study. Their contribution was invaluable and without it this research would not have been possible. Special thanks go to those who devoted their time and energy to gathering and analysing the data to ensure its accuracy and validity.

Funding Information

The authors of this study have not received any funding for their research.

Author's Contributions

Evaristus Didik Madyatmadja: Lead research project, coordinated developer, did experiments, was an instructor, did data analysis and wrote the manuscript.

Aldi: Advised research project, designed the experiment, data analysis, and written manuscript.

Fiona Fheren: Advise data analysis and written manuscript.

Hanny Juwitasary: Advise research project, design the research methodology, data analysis, written manuscript, proofreader.

Helen Angelica and David Jumpa Malem Sembiring: Advise research project, design the research methodology, data analysis, written manuscript and proofreader.

Ethics

The authors confirm that this manuscript has not been published elsewhere and that no ethical issues are involved.

References

Ardiansyah, A. (2017). Program Reminder Meggunakan Sms Gateway Berbasis Android Jakarta, *Repository. Bsi. Ac. Id*,.

Ardianto, R., Rivanie, T., Alkhalifi, Y., Nugraha, F. S., & Gata, W. (2020). Sentiment analysis on E-sports for education curriculum using naive Bayes and support vector machine. *Jurnal Ilmu Komputer dan Informasi*, *13*(2), 109-122. https://doi.org/10.21609/jiki.v13i2.885

Diale, M., Van Der Walt, C., Celik, T., & Modupe, A. (2016, November). Feature selection and support vector machine hyper-parameter optimization for spam detection. In *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)* (pp. 1-7). IEEE. https://doi.org/10.1109/RoboMech.2016.7813162

Golzar, Jawad Noor, Shagofah & Tajik, Omid. (2022). Simple Random Sampling. *International Journal of Eductaion and Language Studies*, 1. 78-82. https://doi.org//10.22034/ijels.2022.162982.

Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, *17*, 26-32. https://doi.org/10.1016/j.procs.2013.05.005

Kalcheva, N., Karova, M., & Penev, I. (2020, September). Comparison of the accuracy of SVM kemel functions in text classification. In *2020 International Conference on Biomedical Innovations and Applications (BIA)* (pp. 141-145). IEEE. https://doi.org/10.1109/BIA50171.2020.9244278

Kavitha, G., & Elango, N. M. (2017). An overview of data mining techniques and its applications. *International Journal of Civil Engineering and Technology*, *8*(12), 1013-1020.

Kusumawati, R., D'arofah, A., & Pramana, P. A. (2019, October). Comparison performance of naive bayes classifier and support vector machine algorithm for twitter's classification of tokopedia services. In *Journal of Physics: Conference Series* (Vol. *1320*, No. 1, p. 012016). IOP Publishing. https://doi.org/10.1088/1742-6596/1320/1/012016

Luo, X. (2021). Efficient English text classification using selected machine learning techniques. *Alexandria Engineering Journal*, *60*(3), 3401-3409. https://doi.org/10.1016/j.aej.2021.02.009

Madyatmadja, E. D., Shinta, Susanti, D., Anggreani, F. & Sembiring, D. J. M. (2022). Sentiment Analysis on User Reviews of Mutual Fund Applications. *Journal of Computer Science*, *18*(10), 885-895. https://doi.org/10.3844/jcssp.2022.885.895

Miao, F., Zhang, P., Jin, L., & Wu, H. (2018, August). Chinese news text classification based on machine learning algorithm. In *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)* (Vol. 2, pp. 48-51). IEEE. 10.1109/IHMSC.2018.10117

Pandiangan, N., Buono, M. L. C., & Loppies, S. H. D. (2020, July). Implementation of decision tree and Naïve Bayes classification method for predicting study period. In *Journal of Physics: Conference Series* (Vol. *1569*, No. 2, p. 022022). IOP Publishing. https://doi.org/10.1088/1742-6596/1569/2/022022

Prasetijo, A. B., Isnanto, R. R., Eridani, D., Soetrisno, Y. A. A., Arfan, M., & Sofwan, A. (2017, October). Hoax detection system on Indonesian news sites based on text classification using SVM and SGD. In *2017 4th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)* (pp. 45-49). IEEE. https://doi.org/10.1109/ICITACEE.2017.8257673

Rianto, Mutiara, A. B., Wibowo, E. P., & Santosa, P. I. (2021). Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation. *Journal of Big Data*, *8*, 1-16. https://doi.org/10.1186/s40537-021-00413-1

Ridzuan, F., & Zainon, W. M. N. W. (2019). A review on data cleansing methods for big data. *Procedia Computer Science*, *161*, 731-738. https://doi.org/10.1016/j.procs.2019.11.177

Sharma, G. (2017). Pros and cons of different sampling techniques. *International Journal of Applied Research*, *3*(7), 749-752. https://www.allresearchjournal.com/archives/2017/vol3issue7/PartK/3-7-69-542.pdf

Sravya, G. S., & Pradeepini, G. (2020). Mobile SMS spam filter techniques using machine learning techniques. *9*, 384-389.

Suntoro, J. (2019). *Data Mining: Algoritma dan Implementasi dengan Pemrograman php*. Elex Media Komputindo. ISBN: 10-6020498816.

Tubagus, M. (2018). Model Pengembangan Short Massage Service (SMS) pada jaringan Seluler. *Potret Pemikiran*, *22*(2). https://doi.org/10.30984/pp.v22i2.783

Zhao, W. (2020, November). Classification of Customer Reviews on E-commerce Platforms Based on Naive Bayesian Algorithm and Support Vector Machine. *In Journal of Physics: Conference Series* (Vol. *1678*, No. 1, p. 012081). IOP Publishing. https://doi.org/10.1088/1742-6596/1678/1/012081