Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 9 No. 1 2024, pp. 74 – 87

Classification Email Spam using Naive Bayes Algorithm and Chi-Squared Feature Selection

Maylinna Rahayu Ningsih*[1], Jumanto[2]
*[1,2]*Computer Science Department, Universitas Negeri Semarang, Semarang, Indonesia*
[3]*Faculty of Computer Science and Information Technology, Universiti Tun Hussein onn Malaysia, Johor, Malaysia*
[4] *Faculty Technology Management, Universiti Tun Hussein Onn Malaysia, Johor, Malaysia.*
E-mail : maylinnarahayuningsih@students.unnes.ac.id*[*1]*, jumanto@mail.unnes.ac.id[2],
**Corresponding author*

Habib al Farih[3], Much Aziz Muslim[4]
[3]*Faculty of Computer Science and Information Technology, Universiti Tun Hussein onn Malaysia, Johor, Malaysia*
[4] *Faculty Technology Management, Universiti Tun Hussein Onn Malaysia, Johor, Malaysia.*
E-mail : bi220041@student.uthm.edu.my[3] , 212muslim@yahoo.com [4]

Received 8 December 2015; Revised 10 February 2016; Accepted 2 March 2016

Abstract - Spam email is a problem that disturbs and harms the recipient. Machine learning is widely used in overcoming email spam because of its ability to classify emails into spam or non-spam. In this research, the Naïve Bayes algorithm is initiated with the Chi-Squared selection feature to classify spam emails. So that the implementation is able to increase accuracy for better performance in classification. The feature selection method is used to direct the model's attention to features that are related to the target variable. In this study, the chi squared feature uses a value of K = 2500, with an accuracy of 98.83% which shows an increase in model performance compared to previous research. So that the Naïve Bayes model with the Chi-Squared selection feature is proven to provide better performance.

Keywords - Email Spam; Naïve Bayes; Chi Squared; Classification; Feature Selection

1. INTRODUCTION

Nowadays, the internet and social media have become an important part of a person's daily life. Email is one of the media used in exchanging information, but it is not free from problems, spam email is a daily problem that disturbs the daily life of people [1], [2], [3]. Spam email is detrimental because it can spread viruses/malware, steal important information, consume bandwidth and things that are personally harmful [4], [5]. On the other hand, the rise of e-commerce companies that use email for advertising has led to an increase in unwanted and indiscriminate bulk emails [6].

The effects of such bulk emails are detrimental, hence the need for action in dealing with them. Machine Learning is an advancement that can help understand email spam. Machine learning has been widely applied in categorization and detection [7], [8], [9][10]. In the case of bulk email, it can be overcome by detection and categorization of spam into spam email and not. Spam detection is an important data analysis [11] for classification of intrusive email messages. But it's not easy to identify whether a message is spam [12].

Machine Learning requires algorithms, in supporting the performance of the algorithms used, we can use various parameters and considerations, one of which is the selection feature. Although selection features can have an impact on text classification, not all of them are beneficial [13]. This feature selection has been done in many text classification studies [14], [15], [16]. One of the selection features is Chi-Squared, which is used to test two events [17]. On [18] chi square is used in dimensional feature reduction and focuses on the features that are needed, so that redundant features can be removed. Chi-Squared will measure the dependency of category variables, in this case spam and non-spam and then applied in feature selection. Testing dependencies between categorical variables, such as spam and non-spam, is easier with Chi-Squared in feature selection. This method helps find and eliminate unnecessary features, improves efficiency, and ensures that the features retained in the model are relevant.

Looking at comparisons in previous research studies has been discussed in almost the same context, such as in the article [19] The ensemble learning technique is utilized in identifying spam on Short Message Service (SMS) spam and email spam, and the classification results utilizing voting are compared, with Decision tree, Multinomial Naive Bayes, Bernoulli Naive Bayes, and Gaussian Naive Bayes being the most used. The email spam dataset has a high accuracy of 92.354% when utilizing Multinomial Naive Bayes, Bernoulli Naive Bayes, and Decision Tree classifications. Similarly, in the process of identifying spam, research [20] discussed email spam identification using Isolation Forest, DBSCAN and feature selection such as chi square to improve accuracy, the results showed 100% accuracy in machine learning implementation and 99% in deep learning implementation. Other considerations in the identification process can be made by paying attention to the context of the text, this is in line with the study [21] which uses the bert-base-cased transformer model with the use of the attention layer to retrieve text connections then the results are compared with BiLSTM (bidirectional Long Short Term Memory) which is a layer of DNN (deep neural network). Memory) which is a layer of DNN (deep neural network), the classification results reached 98.67% accuracy. Other model approaches such as in research [22] use semantic models as an approach by considering the semantics of the word other research. The feature selection used a reduction technique, the result was that Naïve Bayes combined with semantic relationships and words got 94% accuracy, but classification with Support Vector Machine (SVM) got 93% accuracy, then rose to 94% when feature selection was used. Feature selection was shown to improve the accuracy of some classifications compared to using only semantics. Furthermore, the use of binary models is used in research [23] using a binary model by classifying email spam into categories, the results of using SVM with Term Frequency - Inverse Document Frequency (TF-IDF) got an accuracy of 95.53%. While naïve bayes gets the fastest classification time when combined with TF-IDF. Other research related to feature selection [24] proposed the use of sine–cosine algorithm (SCA), a method used to select optimal features in ANN training with an accuracy result of 97.92%.

From the shortcomings of existing research, we conducted this research as an effort to overcome the problem of feature selection that has not achieved high accuracy in email spam detection. As a result, in this study, we employed feature selection to improve the previously unoptimized performance of the Nave Bayes method. Where the Naïve Bayes model is supported by the chi-squared feature selection technique. Naïve Bayes itself has a great influence on text classification, but its performance can be improved by feature selection. So that the selection of chi-square improves the ability of naïve bayes in email spam classification better to increase accuracy and minimize the risk of overfitting.

2. RESEARCH METHOD

The process in this research includes Pre-Processing Stage, Feature Extraction with CountVectorizer, feature selection with Chi-Square and Naïve Bayes modeling. The proposed flowchart is described in Figure 1**.** Then the process explanation will be explained in detail in the next discussion.

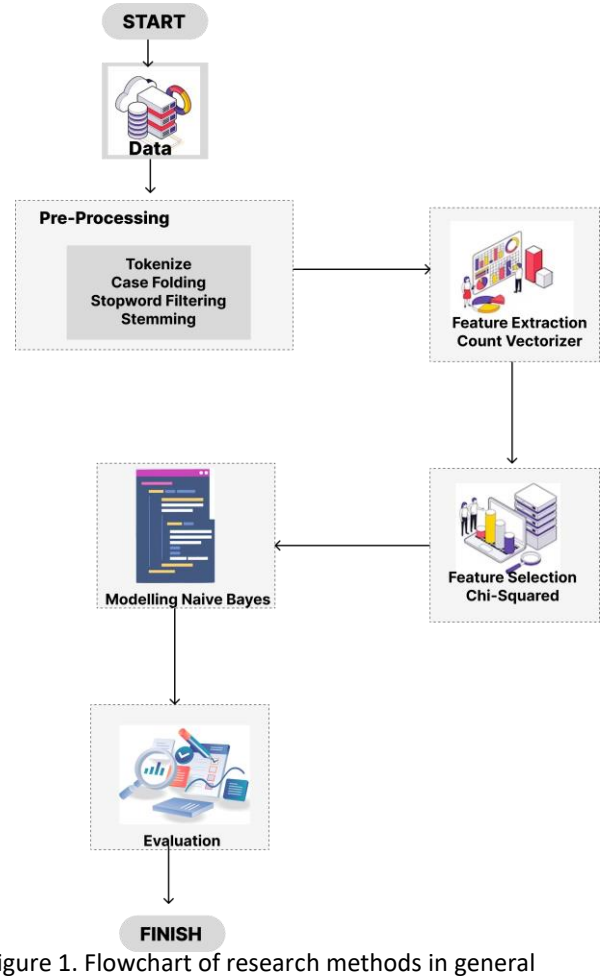

Figure 1. Flowchart of research methods in general

2.1. Data Collection

The dataset used in this research is taken from Kaggle which contains two columns, Category and Message then with spam and ham labels https://www.kaggle.com/datasets/mfaisalqureshi/spam-email . The dataset contains 5558 emails, of which 87% are labeled as ham and 13% are spam. Some examples of sample content datasets can be seen in Table 1 and the proportion of spam and non-spam labels is shown in Figure 2.

Table 1. Sample dataset contents

Category	Message
Ham	Go until jurong point, crazy. Available only ...
Ham	Ok lar... Joking wif u oni...
Spam	Free entry in 2 a wkly comp to win FA Cup fina...
Ham	U dun say so early hor... U c already then say...
Spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, Â£1.50 to rcv
Ham	I HAVE A DATE ON SUNDAY WITH WILL!!
Spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJKGIGHJJGCBL
Ham	Is that seriously how you spell his name?
Spam	Congrats! 1 year special cinema pass for 2 is yours. call 09061209465 now! C Suprman V, Matrix3, StarWars3, etc all 4 FREE! bx420-ip4-5we. 150pm. Dont miss out!

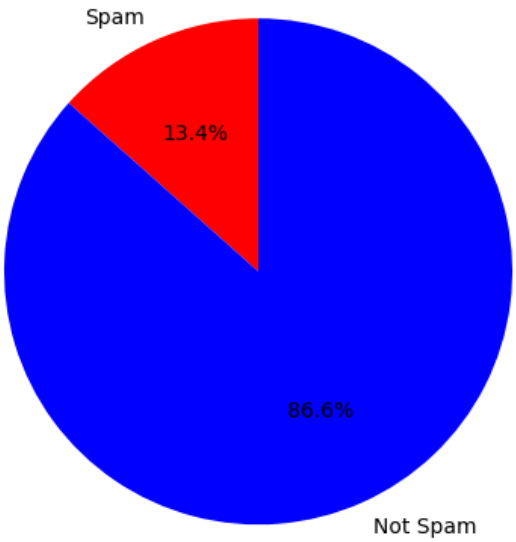

Figure 2. Proportion of Spam vc Not Spam

The amount of spam at 13.4% in the dataset shows that, although spam is not dominant, it is important to continuously improve detection and filtration strategies to ensure better data quality. With a spam note of 86.6%, most of the data is relevant and useful. To guarantee the accuracy and reliability of the analysis performed, efforts to keep this dataset clean should be continuously improved.

2.2. Pre-Processing

Before the dataset continues to the next process, the dataset is cleaned so that the data can be processed in the next analysis. This also aims to improve data quality when spam detection provides better accuracy [25]. The stages carried out are eliminating unnecessary characters and numbers, Tokenizing Text, Case Folding, Stop words Filtering from English Stop words, and Stemming.

2.2.1 Tokenize text

At this stage all text data is converted by separating into words called tokens [26]. It aims to make text data processing easier overall for further processing. It is important to consider rules such as structural equivalence, form parallelism, word economy, reasoning accuracy, assertiveness, coherence, and language logic when creating successful sentences. This stage includes data processed in three different forms. First, the document text is converted into word counts; second, the data is cleaned and filtered; and finally, the document is broken down into words or tokens [27]. The following examples illustrate email messages before and after tokenization:

Before tokenize text
Is that seriously how you spell his name

After tokenize text
Is, that, seriously, how, you, spell, his, name

2.2.2 Case folding (convert to lowercase)

The following examples illustrate email messages before and after tokenization: This process converts the text to all lowercase letters [12]. In this process, all characters from the letters "A" to "Z" present in the data are converted into the letters "a" to "z". This process also eliminates non-word characters such as numbers, symbols, and punctuation marks, so that the remaining text is only alphabetical from a to z. This needs to be done, because the form of words and sentences in email spam has a different diversity, so for a good process in analyzing converting to lowercase needs to be done. An example of case folding results is shown in Table 2.

Table 1. Sample dataset contents

Example Teks	Case Folding Result
Do you want a new video handset? 750 anytime any network mins? Half Price Line Rental? Camcorder? Reply or call 08000930705 for delivery tomorrow	do you want a new video handset? 750 anytime any network mins? half price line rental? camcorder? reply or call 08000930705 for delivery tomorrow

2.2.3 Stopword filtering

This step removes frequently occurring words or terms such as pronouns, prepositions and those that do not provide important information and hinder the process [28]. In other words, this stopword removal process does not negatively impact the model we are training for our task as it removes low-level information from our text and gives focus to more important information. These low-quality topics learn a background distribution for stopwords, but words without infrequent content may inadvertently correlate with content-packed topic terms, while words like "the" are so frequent that they remain prominent in many topics. Since they are rarely used, the preceding terms need not interfere. This process also aims at reducing the size of the dataset, which in turn reduces the number of tokens required for training. The Stopword removal process is utilized in the hope that it will help learn a highquality language model. The Stop words used in this research are 'English' Stop words.

2.2.3 Stemming

Stemming converts words into recognizable similar structures, finding the root word of the word by removing affixes [28]. After that, this basic word form will be stored and processed [29]. This process is done to make the data analysis process easier by reducing the inflectional or derivative word to its base form.

2.3. Feature extraction using CountVectorizer

Feature extraction in this research uses CountVectorizer on training data and test data. Where CountVectorizer here is used to process email message text and then convert email text into numeric vectors so that it can be used in the model. It is used to convert a given text into a vector that is based on the frequency, or count, of each word that appears throughout the text [30]. This is especially helpful when you have a lot of text, such as email messages, where you need to convert each word into a vector to use for further text analysis.

2.4. Chi-Squared Feature Selection

Chi-Squared feature selection is a statistical technique/method to look at variable and categorial relationships [31]. This feature selection helps in determining the important/relevant features in email spam. The parameter value K = 2500 is used, as the best available feature selection. Chi-Square calculation can be describe in Equation (1) and Equation (2).

$$X^2(t_i, C_k) = \frac{N\ x\ (ad-bc)^2}{(a+c)x(b+d)x(a+b)x(c+d)}\ [17] \qquad (1)$$

The calculation in [17] explains that "a" represents the number of records/instances in category Ck that contain the term ti, "b" represents the number of records/instances that are not in category Ck and do not contain the term ti, "c" represents the number of records/instances in category Ck that do not contain the term ti, and "d" represents the number of records/instances that are not in category Ck. N is the whole document that was used. Equation computes the chi square value for each phrase.

Then in this study the calculation is developed in outline, which is written in Equation (2)

$$Chi\ Square = \frac{(Observed-Excepted)^2}{Excepted} \qquad (2)$$

Where the Observed value is the contingency table (number of observations in each cell) and the Expected value is the expected value based on the null hypothesis assumption.

2.5. Multinomial Naïve Bayes (MNB)

Naïve Bayes is a simple classification that calculates combinations of values and frequencies in a dataset [32], [33], [34]. Before using the model, the data is separated with a percentage of 80% training data and 20% testing data. Naïve Bayes performs classification with similar capabilities to neural networks and decision trees, one of the simplest classifications but provides high accuracy and speed [26], [35]. Multinomial Naïve Bayes is a Nave Bayes extension that uses multinomial distribution along with the number of times a word appears in a text to tackle text classification problems [36]. In this research, Naïve Bayes classification is used to identify spam and non-spam from email datasets. Naïve Bayes uses the Bayes theorem where the calculation is in the formula (3) & (4) below.

$$P\ (A|B) = \frac{P(B|A)xP(A)}{P(B)}\ [37] \qquad (3)$$

$$P\ (spam|word) = \frac{P(word|spam)xP(spam)}{P(spam).\ P(word|spam)+p(non-spam).p(word|non-spam)} \qquad [32] \qquad (4)$$

2.5. Measurement Model

Predictions and the actual condition of the data produced by machine learning algorithms can both be displayed using Confusion Matrix. False Positives (FP) and False Negatives (FN) are the correct number of positive classes, the number of false positive classes, the correct negative class, and the incorrect negative class on the data, respectively, in the Confusion Matrix [38]. Accuracy, Recall, Precision and F1 score can be calculated using this matrix. The percentage of samples that were correctly assigned to a group determines accuracy. The total sample size for the test dataset [39], as stated in Equation (5).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Recall is a ratio of all true positive data, including TP and FN, to true positive forecasts. The formula for calculating Recall is Equation (6).

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

If precision measures the match between parts of the data taken with the information needed. precision calculations are shown in the Equation (7).

$$Precision = \frac{TP}{TP+FP} \quad 7)$$

The F1 score, also known as the average recall value and precision, measures a researcher's capacity to identify all positive samples of data and to avoid mislabeling negative classes as positive ones. The [0, 1] range of values are accepted for the F1 score. The equation used to determine the f1 score is Equation (8).

$$\frac{2\ x\ precision\ x\ recall}{precision+recall} \quad (8)$$

Broadly speaking, four performance metrics are used to evaluate experiments with usability as shown in Table 3.

Table 3. Brief concept of performance metrics

Evaluation Metrics	Explanation
Accuracy	The number of incidents accurately categorised
Precision	The percentage of relevant cases among the retrieved instances
Recall	A percentage of the total number of relevant instances actually retrieved
F1 score	Precision and recall harmonic mean

3. RESULTS AND DISCUSSION

In producing good data analysis, the pre-processing process plays an important role, the following examples of pre-processing data results are shown in Table 4.

Table 4 Results Pre-Processing

Before Pre-Processing	After Pre-Processing
Do 1 thing! Change that sentence into: "Because i want 2 concentrate in my educational career im leaving here..	thing change sentence want concentr educ career im leave
Free Msg: get Gnarls Barkleys "Crazy" ringtone TOTALLY FREE just reply GO to this message right now!	free msg get gnarl barkley crazi rington total free repli go messag right
reat NEW Offer - DOUBLE Mins & DOUBLE Txt on best Orange	reat new offer doubl min doubl txt best orang tariff get latest

tariffs AND get latest camera phones 4 FREE! Call MobileUpd8 free on 08000839402 NOW! or 2stoptxt T&Cs	camera phone free call mobileupd free stoptxt cs
We'll you pay over like <#> yrs so its not too difficult	pay like lt gt yr difficult
IM FINE BABES AINT BEEN UP 2 MUCH THO! SAW SCARY MOVIE YEST ITS QUITE FUNNY! WANT 2MRW AFTERNOON? AT TOWN OR MALL OR SUMTHIN?xx	im fine babe aint much tho saw scari movi yest quit funni want mrw afternoon town mall sumthin xx
YOU ARE CHOSEN TO RECEIVE A £350 AWARD! Pls call claim number 09066364311 to collect your award which you are selected to receive as a valued mobile customer.	chosen receiv award pl call claim number collect award select receiv valu mobil custom

Furthermore, in visualizing the results of feature extraction, PCA (Principal component analysis) is used.This method puts multidimensional data into a small space and describes the research object as a whole in the form of several main components. Principal component analysis (PCA) processes data with principal component eigenvalues [40]. The following are the results of PCA on training data and testing data in Figure 3.

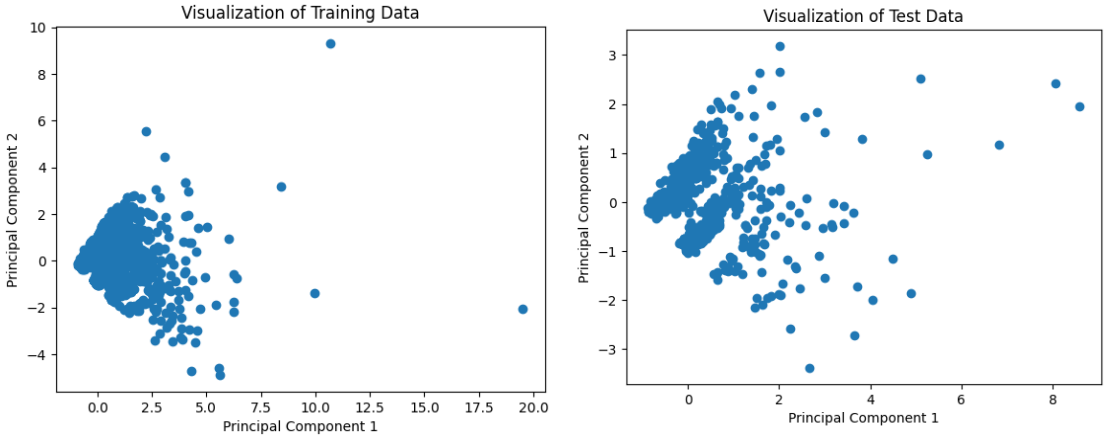

Figure 3. PCA Result

There is a difference between the training data and the test data due to higher variation. This could indicate that the test data includes more varied examples or greater variation compared to the training data. When the data is ready for processing, it is divided into training and testing data, with a weight of 80% for training and 20% for testing data. Then the CountVectorizer feature extraction and Chi-Squared feature selection are used with a value of K = 2500, this value is the best K value after being tested on several K values such as Chi-Squared Scores of Features in Figure 4.

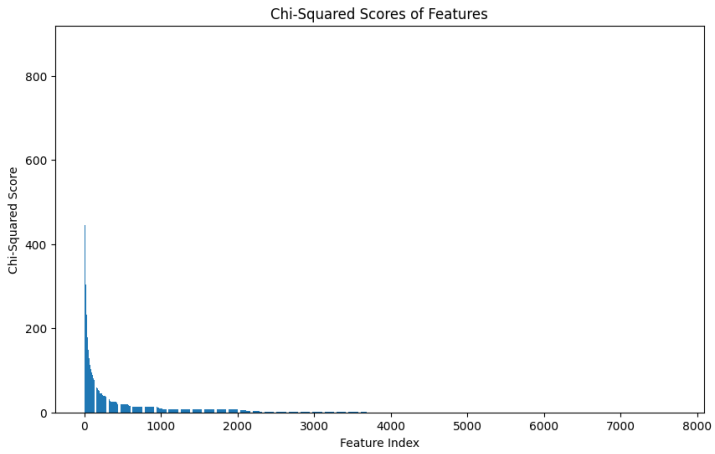

Figure 4. Chi-Squared Scores of Features

The Multinomial Nave Bayes model is then used to evaluate the model and calculate the evaluation matrix, the results are discussed in Table 5.

Table 5. Evaluation metrics table

Evaluation Metrics	Value
Accuracy	98.83%
Precision	98.83%
Recall	98.83%
F1-score	98.83%

The results shown have a high performance for each evaluation metric which is 98.83%. The evaluation value of the metric is calculated by comparing the test results from the Confusion Matrix results. In fact, Confusion Matrix is also the most widely used [41]. Classification algorithms can be compared and evaluated for performance by looking at the Confusion Matrix [42]. To see the performance of Naïve Bayes classification this study involves Confusion Matrix, the results are in Figure 5.

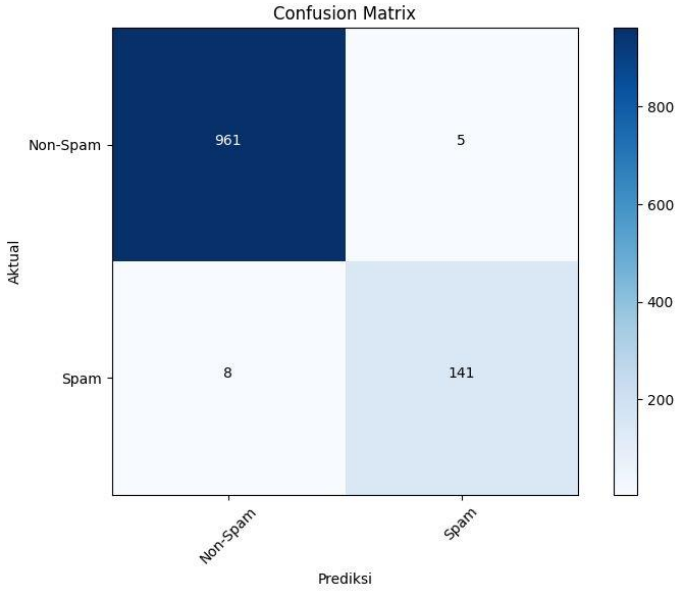

Figure 5. Confusion matrix calculation results

The Confusion Matrix findings compare Nave Bayes categorization and prediction. The following values are generated: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) [43]. In this study, the number of correct predictions as spam or TP values was 141 emails correctly predicted as spam. While those that were wrongly predicted (FP) there were 5 emails that were wrongly predicted as spam by being NonSpam. For Non-Spam TN there are 961 emails correctly predicted as non-spam. The remaining FN value is 8 emails that should be a spa wrongly predicted as non-spam. This means that the model's performance in classifying is high and the classification error rate is small. The diagonal elements in a confusion matrix indicate examples that were correctly classified by the model, while the off-diagonal elements indicate examples that were classified incorrectly. The ROC curve can be calculated using the confusion matrix [41] as shown in Figure 6.

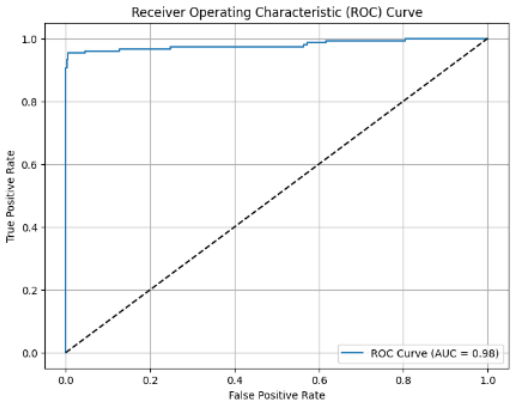

Figure 6. ROC results

Furthermore, to see the words that are often used as spam in emails, such as the words "Free", "now", "call", "call now", and other words can be seen in Figure 7.

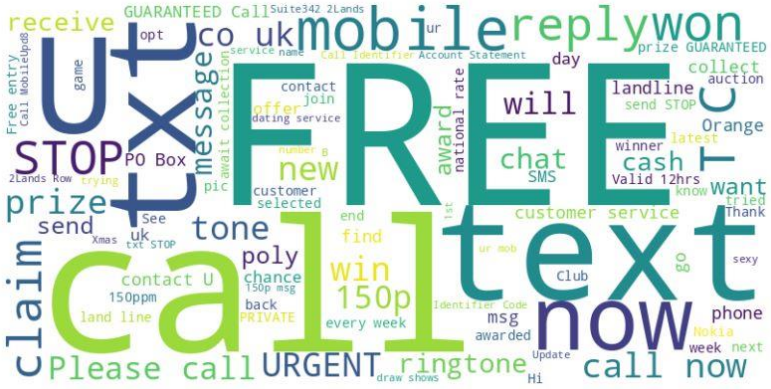

Figure 7. Word Cloud of spam words

To determine the comparison of the resulting model that successfully provides good performance, a comparison with previous research is carried out as in Table 6.

Table 6 - Model performance comparison

Comparison	Algorithm	Accuracy (%)
In [19]	Gaussian Naive Bayes	92.354
In [21]	BiLSTM + DNN	98.67
In [22]	Feature selection (CFS) + Semantic relations and similarity measures	94
In [23]	SVM + TF -DF	95.53
In [24]	ANN	97.92
In [44]	Naïve Bayes + TF-IDF	98.5
In [45]	Auto-GA-RWN and GA-kNN	96.7
In [46]	Naive Bayes + Bigrams	93
Proposed method	**Naïve Bayes + Chi Square**	**98,83**

Table 6 shows a comparison of the performance of the created model and models from previous research. The experimental results show that the proposed method, which uses Naive Bayes with Chi-Square as the feature selection method, achieves the highest accuracy of 98.83%. This model is compared with other models such as Gaussian Naive Bayes, BiLSTM + DNN, feature selection (CFS) + semantic relationship and similarity metrics, SVM + TF-DF, ANN, Naive Bayes + TF-IDF, Auto-GA-RWN and GA-kNN, and Naive Bayes + Bigrams.

In Naïve Bayes models, using Chi-Square as a feature selection method significantly improves accuracy and helps to identify and retain the most relevant features; if not used, the model may face the risk of containing redundant or less informative features, which may reduce accuracy and efficiency. Therefore, adding Chi-Square into feature selection is essential to improve the performance of spam detection models.

4. CONCLUSION

In this study, the Naïve Bayes model combined with Count Vectorizer feature extraction and Chi-Squared feature selection with K=2500 managed to provide better performance. The findings reveal that the performance for Accuracy, Precision, Recall, and F1 Score is at 98.83%. The use of Chi-Squared with K=2500 is the best value in improving model performance. This feature selection helps in determining important/relevant features in email spam. Although the model still has errors in predicting classification, it is good enough and improved from the previous article. The shortcomings of this model may later be improved with other classification models.

REFERENCES

[1] R. M. A. Mohammad, "Applied Computing and Informatics A lifelong spam emails classification model," *Appl. Comput. Informatics*, no. xxxx, hal. 1–10, 2020, doi: 10.1016/j.aci.2020.01.002.

[2] T. A. Almeida dan J. Almeida, "Spam filtering : how the dimensionality reduction affects the accuracy of Naive Bayes classifiers," hal. 183–200, 2011, doi: 10.1007/s13174-010-0014-7.

[3] Y. Cohen, D. Hendler, dan A. Rubin, "US CR," *Knowledge-Based Syst.*, 2017, doi: 10.1016/j.knosys.2017.11.011.

[4] T. Gangavarapu dan C. D. J. B. Chanduka, *Applicability of machine learning in spam and phishing email filtering : review and approaches*, vol. 53, no. 7. Springer Netherlands, 2020.

[5] U. Murugavel dan R. Santhi, "Materials Today : Proceedings Detection of spam and threads identification in E-mail spam corpus using content based text analytics method," *Mater. Today Proc.*, no. xxxx, 2020, doi: 10.1016/j.matpr.2020.04.742.

[6] B. Kagan dan B. Akay, "Spam filtering using a logistic regression model trained by an artificial bee colony algorithm," *Appl. Soft Comput. J.*, vol. 91, hal. 106229, 2020, doi: 10.1016/j.asoc.2020.106229.

[7] N. Hidayat dan M. F. Al Hakim, "Halal Food Restaurant Classification Based on Restaurant Review in Indonesian Language Using Machine Learning," vol. 8, no. 2, hal. 314–319, 2021, doi: 10.15294/sji.v8i2.25356.

[8] H. Fang, J. Xiao, dan Y. Wang, "International Journal of Electrical Power and Energy Systems A machine learning-based detection framework against intermittent electricity theft attack," *Int. J. Electr. Power Energy Syst.*, vol. 150, no. March, hal. 109075, 2023, doi: 10.1016/j.ijepes.2023.109075.

[9] S. Schulz, M. Becker, M. R. Groseclose, S. Schadt, dan C. Hopf, "Advanced MALDI mass spectrometry imaging in pharmaceutical research and drug development," *Curr. Opin. Biotechnol.*, vol. 55, hal. 51–59, 2019, doi: 10.1016/j.copbio.2018.08.003.

[10] M. Schulz dan T. Schr, "Monitoring machine learning models : a categorization of challenges and methods," vol. 5, no. July, hal. 105–116, 2022, doi:

10.1016/j.dsm.2022.07.004.
[11] S. Rahman, “An efficient hybrid system for anomaly detection in social networks,” 2021.
[12] J. Fattahi, “SpaML : a Bimodal Ensemble Learning Spam Detector based on NLP Techniques,” no. Ml, hal. 107–112, 2021.
[13] W. Binsaeedan dan S. Alramlawi, “Knowledge-Based Systems CS-BPSO : Hybrid feature selection based on chi-square and binary PSO algorithm for Arabic email authorship analysis,” *Knowledge-Based Syst.*, vol. 227, hal. 107224, 2021, doi: 10.1016/j.knosys.2021.107224.
[14] G. Kou, P. Yang, Y. Peng, F. Xiao, Y. Chen, dan F. E. Alsaadi, “Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods,” *Appl. Soft Comput. J.*, vol. 86, hal. 105836, 2020, doi: 10.1016/j.asoc.2019.105836.
[15] R. Cekik dan A. K. Uysal, “Expert Systems with Applications A novel filter feature selection method using rough set for short text data,” *Expert Syst. Appl.*, vol. 160, hal. 113691, 2020, doi: 10.1016/j.eswa.2020.113691.
[16] K. Thirumoorthy, “Optimal feature subset selection using hybrid binary Jaya optimization algorithm for text classification,” *Sādhanā*, vol. 45, no. 1, hal. 1–13, 2020, doi: 10.1007/s12046-020-01443-w.
[17] U. I. Larasati, M. A. Muslim, dan R. Arifudin, “Improve the Accuracy of Support Vector Machine Using Chi Square Statistic and Term Frequency Inverse Document Frequency on Movie Review Sentiment Analysis,” vol. 6, no. 1, hal. 138–149, 2019.
[18] L. Allen, C. Ahakonye, C. I. Nwakanma, J. Lee, dan D. Kim, “Internet of Things SCADA intrusion detection scheme exploiting the fusion of modified decision tree and Chi-square feature selection,” *Internet of Things*, vol. 21, no. September 2022, hal. 100676, 2023, doi: 10.1016/j.iot.2022.100676.
[19] V. Gupta, A. Mehta, A. Goel, U. Dixit, dan A. C. Pandey, *Learning*. Springer Singapore.
[20] F. Hossain, “Analysis of Optimized Machine Learning and Deep Learning Techniques for Spam Detection,” 2021.
[21] H. A. M. Bert, “ScienceDirect Spam Spam Email Email Detection Detection Using Using Deep Deep Learning Learning Techniques Techniques,” *Procedia Comput. Sci.*, vol. 184, no. 2019, hal. 853–858, 2021, doi: 10.1016/j.procs.2021.03.107.
[22] E. M. Bahgat, S. Rady, W. Gad, dan I. F. Moawad, “Efficient email classification approach based on semantic methods,” *Ain Shams Eng. J.*, vol. 9, no. 4, hal. 3259–3269, 2018, doi: 10.1016/j.asej.2018.06.001.
[23] J. Velasco-mata, “Classification of Spam Emails through Hierarchical Clustering and Supervised Learning.”
[24] R. Talaei, P. Yaser, dan R. Mohsen, “Spam detection through feature selection using artificial neural network and sine – cosine algorithm,” *Math. Sci.*, vol. 14, no. 3, hal. 193–199, 2020, doi: 10.1007/s40096-020-00327-8.
[25] K. Taghandiki, “Building an Effective Email Spam Classification Model with spaCy,” hal. 1–5.
[26] S. Ernawati, “Implementation of The Naïve Bayes Algorithm with Feature Selection using Genetic Algorithm for Sentiment Review Analysis of Fashion Online Companies,” *2018 6th Int. Conf. Cyber IT Serv. Manag.*, no. Citsm, hal. 1–5, 2018, doi: 10.1109/CITSM.2018.8674286.
[27] N. Parveen, P. Chakrabarti, B. T. Hung, dan A. Shaik, “Twitter sentiment analysis using hybrid gated attention recurrent network,” *J. Big Data*, 2023, doi: 10.1186/s40537-023-00726-3.
[28] S. Suryawanshi, “Email Spam Detection : An Empirical Comparative Study of Different

ML and Ensemble Classifiers," hal. 69–74, 2019.
[29] P. Widyaningrum, Y. Ruldeviyani, R. Dharayani, P. Widyaningrum, Y. Ruldeviyani, dan R. Dharayani, "ScienceDirect ScienceDirect Sentiment Analysis to Assess the Community ' s Enthusiasm Sentiment Analysis to Assess the Community ' s Enthusiasm Towards the Development Chatbot Using an Appraisal Theory Towards the Development Chatbot Using an Appraisal Theory," *Procedia Comput. Sci.*, vol. 161, hal. 723–730, 2019, doi: 10.1016/j.procs.2019.11.176.
[30] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, dan H. Al Najada, "Survey of review spam detection using machine learning techniques," *J. Big Data*, 2015, doi: 10.1186/s40537-015-0029-9.
[31] N. Peker dan C. Kubat, "Application of Chi-square discretization algorithms to ensemble classification methods," *Expert Syst. Appl.*, vol. 185, no. June, hal. 115540, 2021, doi: 10.1016/j.eswa.2021.115540.
[32] N. F. Rusland, N. Wahid, dan S. Kasim, "Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets Analysis of Na ¨ ıve Bayes Algorithm for Email Spam Filtering across Multiple Datasets," doi: 10.1088/1757-899X/226/1/012091.
[33] I. P. Wardhani, Y. I. Chandra, dan F. Yusra, "Application of the Naïve Bayes Classifier Algorithm to Analyze Sentiment for the Covid-19 Vaccine on Twitter in Jakarta," vol. 07, no. 01, hal. 1–18, 2023.
[34] V. A. Fitri, R. Andreswari, M. A. Hasibuan, V. A. Fitri, R. Andreswari, dan M. A. Hasibuan, "ScienceDirect ScienceDirect ScienceDirect Sentiment Analysis of Social Media Twitter with Case of Anti- Sentiment Analysis of Social Media Twitter with Case of Anti- LGBT Campaign in Indonesia using Naïve Bayes , Decision Tree , LGBT Campaign in Indonesia using Naïve Bayes , Decision Tree , and Random Forest Algorithm and Random Forest Algorithm," *Procedia Comput. Sci.*, vol. 161, hal. 765–772, 2019, doi: 10.1016/j.procs.2019.11.181.
[35] P. Aliandu, "Sentiment Analysis to determine Accommodation , Shopping and Culinary Location on Foursquare in Kupang City," *Procedia - Procedia Comput. Sci.*, vol. 72, hal. 300–305, 2015, doi: 10.1016/j.procs.2015.12.144.
[36] V. Balakrishnan dan W. Kaur, "ScienceDirect ScienceDirect String-based Multinomial Naïve Bayes for Emotion Detection String-based Multinomial Naïve Bayes for Emotion Detection among Facebook Diabetes Community among Facebook Diabetes Community," *Procedia Comput. Sci.*, vol. 159, hal. 30–37, 2019, doi: 10.1016/j.procs.2019.09.157.
[37] D. Van Herwerden, J. W. O. Brien, P. M. Choi, K. V Thomas, P. J. Schoenmakers, dan S. Samanipour, "Chemometrics and Intelligent Laboratory Systems Naive Bayes classi fi cation model for isotopologue detection in LC-HRMS data," *Chemom. Intell. Lab. Syst.*, vol. 223, no. January, hal. 104515, 2022, doi: 10.1016/j.chemolab.2022.104515.
[38] M. Aziz, T. Lailatul, D. Ananda, dan A. Pertiwi, "Intelligent Systems with Applications New model combination meta-learner to improve accuracy prediction P2P lending with stacking ensemble learning *," *Intell. Syst. with Appl.*, vol. 18, no. February, hal. 200204, 2023, doi: 10.1016/j.iswa.2023.200204.
[39] D. Chicco dan G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," hal. 1–13, 2020.
[40] Y. Zhang, G. Wang, X. Wang, H. Fan, B. Shen, dan K. Sun, "Energy Geoscience TOC estimation from logging data using principal component analysis," *Energy Geosci.*, vol. 4, no. 4, hal. 100197, 2023, doi: 10.1016/j.engeos.2023.100197.
[41] A. Luque, A. Carrasco, A. Martín, dan A. De, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern*

Recognit., vol. 91, hal. 216–231, 2019, doi: 10.1016/j.patcog.2019.02.023.

[42] D. Valero-carreras, J. Alcaraz, dan M. Landete, "Computers and Operations Research Comparing two SVM models through different metrics based on the confusion matrix," *Comput. Oper. Res.*, vol. 152, no. April 2022, hal. 106131, 2023, doi: 10.1016/j.cor.2022.106131.

[43] L. P. Lim dan M. M. Singh, "Journal of Information Security and Applications Resolving the imbalance issue in short messaging service spam dataset using cost-sensitive techniques," vol. 54, 2020, doi: 10.1016/j.jisa.2020.102558.

[44] F. Jáñez-martino, R. Alaiz-rodríguez, V. González-castro, E. Fidalgo, dan E. Alegre, "Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach," *Appl. Soft Comput.*, vol. 139, hal. 110226, 2023, doi: 10.1016/j.asoc.2023.110226.

[45] M. Mafarja, M. A. Hassonah, dan H. Fujita, "PT US CR," *Inf. Fusion*, 2018, doi: 10.1016/j.inffus.2018.08.002.

[46] A. Ligthart, C. Catal, dan B. Tekinerdogan, "Analyzing the effectiveness of semi-supervised learning approaches for opinion spam classification," *Appl. Soft Comput. J.*, vol. 101, hal. 107023, 2021, doi: 10.1016/j.asoc.2020.107023.