ISSN 2548-8368 (media online)

Jurnal Media Informatika Budidarma

MIB

STMIK Budi Darma Journal

Diterbitkan Oleh :

STMIK Budi Darma Medan

Jl. Sisingamangaraja No.338 Simpang Limun Medan
Telp. 061-7875998
http://www.stmik-budidarma.ac.id

Jurnal Media Informatika Budidarma	Volume : 6 No. 3	Halaman: 1282-1788	Medan Juli 2022	ISSN 2548-8368 (media online)

Vol 6, No 3 (2022): Juli 2022

Table of Contents

Articles

Abdul Fadlil (Universitas Ahmad Dahlan, Yogyakarta, Indonesia)
Imam Riadi (Universitas Ahmad Dahlan, Yogyakarta, Indonesia)

DOI: 10.30865/mib.v6i3.4099 Abstract View 3 times Citations ?

Pengujian ISO 25010 Pada Smart Chair Akupresure Berbasis Internet Of Things (IoT) 1476-1483

Diki Daryanto (STMIK Amik Riau, Riau, Indonesia)
M. Khairul Anam (STMIK Amik Riau, Riau, Indonesia)
Yoyon Efendi (STMIK Amik Riau, Riau, Indonesia)
Rahmaddeni Rahmaddeni (STMIK Amik Riau, Riau, Indonesia)

DOI: 10.30865/mib.v6i3.4134 Abstract View 3 times Citations ?

Sentiment Analysis of Hate Speech on Twitter Public Figures with AdaBoost and XGBoost Methods 1484-1491

Daffa Ulayya Suhendra (Universitas Telkom, Bandung, Indonesia)
Jondri Jondri (Universitas Telkom, Bandung, Indonesia)
Indwiarti Indwiarti (Universitas Telkom, Bandung, Indonesia)

DOI: 10.30865/mib.v6i3.4394 Abstract View 23 times Citations ?

Perancangan Alat Identifikasi Wajah Dengan Algoritma You Only Look Once (YOLO) Untuk Presensi Mahasiswa 1492-1500

Irma Salamah (Politeknik Negeri Sriwijaya, Palembang, Indonesia)
M. Redho Ali Said (Politeknik Negeri Sriwijaya, Palembang, Indonesia)
Sopian Soim (Politeknik Negeri Sriwijaya, Palembang, Indonesia)

DOI: 10.30865/mib.v6i3.4399 Abstract View 6 times Citations ?

Pengaruh Distribusi Panjang Data Teks pada Klasifikasi: Sebuah Studi Awal 1501-1508

Said Al Faraby (Telkom University, Bandung, Indonesia)
Ade Romadhony (Telkom University, Bandung, Indonesia)

DOI: 10.30865/mib.v6i3.4259 Abstract View 6 times Citations ?

Aplikasi Prakiraan Perkembangan Covid-19 Di Indonesia Menggunakan Metode Single Exponential Smoothing Berbasis Web 1509-1516

Tsinmi Tri Azkiya Waslin (Universitas Islam Sumatera Utara, Medan, Indonesia)
Oris Krianto Sulaiman (Universitas Islam Sumatera Utara, Medan, Indonesia)
Tasliyah Haramaini (Universitas Islam Sumatera Utara, Medan, Indonesia)

DOI: 10.30865/mib.v6i3.4408 Abstract View 6 times Citations ?

Penerapan Firebase Realtime Database Pada Aplikasi Media Informasi dan Pendaftaran Training IT Berbasis Android 1517-1525

Angga Arindra Shonta (Universitas AMIKOM Yogyakarta, Yogyakarta, Indonesia)
Laily Nur Hamidah (Universitas AMIKOM Yogyakarta, Yogyakarta, Indonesia)
Muhamad Hasan (Universitas AMIKOM Yogyakarta, Yogyakarta, Indonesia)
Melany Mustika Dewi (Universitas AMIKOM Yogyakarta, Yogyakarta, Indonesia)
Yuli Astuti (Universitas AMIKOM Yogyakarta, Yogyakarta, Indonesia)
Irma Rofni Wulandari (Universitas AMIKOM Yogyakarta, Yogyakarta, Indonesia)

DOI: 10.30865/mib.v6i3.4040 Abstract View 7 times Citations ?

Penerapan Extreme Programming dalam Pengembangan Fitur Interoperabilitas Pada Aplikasi Bioinformatika 1526-1535

Edrian Hadinata (Universitas Harapan Medan, Medan, Indonesia)
Tantri Hidayati Sinaga (Universitas Harapan Medan, Medan, Indonesia)

DOI: 10.30865/mib.v6i3.4238 Abstract View 8 times Citations ?

Penerapan Teknologi Stack MERN pada Aplikasi Service Manajemen Bengkel Berbasis Web 1536-1544

Moch. Akbar Maulana (Universitas Amikom Yogyakarta, Yogyakarta, Indonesia)
Haryoko Haryoko (Universitas Amikom Yogyakarta, Yogyakarta, Indonesia)
Banu Santoso (Universitas Amikom Yogyakarta, Yogyakarta, Indonesia)
Lukman Lukman (Universitas Amikom Yogyakarta, Yogyakarta, Indonesia)

DOI: 10.30865/mib.v6i3.4147 Abstract View 8 times Citations ?

Perbandingan Metode Naïve Bayes dan Support Vector Machine Untuk Analisis Sentimen Terhadap Vaksin Astrazeneca di Twitter 1545-1553

Eva Rahma Indriyani (Institut Teknologi Telkom Purwokerto, Banyumas, Indonesia)
Paradise Paradise (Institut Teknologi Telkom Purwokerto, Banyumas, Indonesia)
Merlinda Wibowo (Institut Teknologi Telkom Purwokerto, Banyumas, Indonesia)

DOI: 10.30865/mib.v6i3.4220 Abstract View 7 times Citations ?

Sentiment Analysis Pada Masyarakat Terhadap LRT Kota Palembang Menggunakan Metode Improved K-Nearest Neighbor 1554-1561

Siti Nur Arafah (Universitas Sriwijaya, Palembang, Indonesia)
Fathoni Fathoni (Universitas Sriwijaya, Palembang, Indonesia)

DOI: 10.30865/mib.v6i3.4434 Abstract View 17 times Citations ?

Implementasi Data Mining dengan Algoritma Naïve Bayes untuk Profiling Korban Penipuan Online di Indonesia 1562-1572

Sunardi Sunardi (Universitas Ahmad Dahlan, Yogyakarta, Indonesia)
Abdul Fadlil (Universitas Ahmad Dahlan, Yogyakarta, Indonesia)
Nur Makkie Perdana Kusuma (Universitas Ahmad Dahlan, Yogyakarta, Indonesia)

DOI: 10.30865/mib.v6i3.3999 Abstract View 3 times Citations ?

Vina Winda Sari (STMIK Triguna Dharma, Medan, Indonesia)
Muhammad Zunaidi (STMIK Triguna Dharma, Medan, Indonesia)
Asyahri Hadi Nasyuha (STMIK Triguna Dharma, Medan, Indonesia)
Marsono Marsono (STMIK Triguna Dharma, Medan, Indonesia)

DOI: 10.30865/mib.v6i3.4140 Abstract View 9 times Citations ?

Penerapan Clustering K-Means untuk Pengelompokan Tingkat Kepuasan Pengguna Lulusan Perguruan Tinggi

1693-1700

Dikky Praseptian M (Universitas Ahmad Dahlan, Yogyakarta, Indonesia)
Abdul Fadlil (Universitas Ahmad Dahlan, Yogyakarta, Indonesia)
Herman Herman (Universitas Ahmad Dahlan, Yogyakarta, Indonesia)

DOI: 10.30865/mib.v6i3.4191 Abstract View 3 times Citations ?

Penerapan metode 7S McKinsey pada Ebay sebagai Strategi E-commerce & Bonus Demography Menghadapi Globalisasi

1701-1711

Hasna Widya Pratiwi (Universitas Diponegoro, Semarang, Indonesia)
Fuad Mas'ud (Universitas Diponegoro, Semarang, Indonesia)

DOI: 10.30865/mib.v6i3.4484 Abstract View 13 times Citations ?

Perbandingan Metode AHP dan TOPSIS untuk Pemilihan Karyawan Berprestasi

1712-1722

Musri Iskandar Nasution (Universitas Ahmad Dahlan, Yogyakarta, Indonesia)
Abdul Fadlil (Universitas Ahmad Dahlan, Yogyakarta, Indonesia)
Sunardi Sunardi (Universitas Ahmad Dahlan, Yogyakarta, Indonesia)

DOI: 10.30865/mib.v6i3.4194 Abstract View 11 times Citations ?

Implementasi XGBoost Pada Keseimbangan Liver Patient Dataset dengan SMOTE dan Hyperparameter Tuning Bayesian Search

1723-1729

Rahmad Ubaidillah (Universitas Lambung Mangkurat, Banjarbaru, Indonesia)
Muliadi Muliadi (Universitas Lambung Mangkurat, Banjarbaru, Indonesia)
Dodon Turianto Nugrahadi (Universitas Lambung Mangkurat, Banjarbaru, Indonesia)
M Reza Faisal (Universitas Lambung Mangkurat, Banjarbaru, Indonesia)
Rudy Herteno (Universitas Lambung Mangkurat, Banjarbaru, Indonesia)

DOI: 10.30865/mib.v6i3.4146 Abstract View 8 times Citations ?

Sistem Pendukung Keputusan Penentuan Profesi Mahasiswa Informatika Menggunakan Metode WP-RIASEC

1730-1739

Raihan Aqila Taufik (Universitas Ahmad Dahlan, Yogyakarta, Indonesia)
Miftahurrahma Rosyda (Universitas Ahmad Dahlan, Yogyakarta, Indonesia)

DOI: 10.30865/mib.v6i3.4312 Abstract View 42 times Citations ?

Penerapan Business Intelligence Terhadap Data Penjualan UMKM (Foodendez) Menggunakan Metode Algoritma Apriori Dalam Menentukan Segmentasi Pasar

1740-1745

Akhmad Rafi Oktavian (Universitas Widyatama, Bandung, Indonesia)
Fitrah Rumaisa (Universitas Widyatama, Bandung, Indonesia)

DOI: 10.30865/mib.v6i3.4338 Abstract View 16 times Citations ?

Penerapan Metode VIKOR (Visekriterijumsko Kompromisno Rangiranje) dalam Pengambilan Keputusan Pemilihan Emulator Android pada Komputer

1746-1755

Renny Pusita Sari (Universitas Tanjungpura, Pontianak, Indonesia)
Meilia Susanti (Universitas Tanjungpura, Pontianak, Indonesia)

DOI: 10.30865/mib.v6i3.4205 Abstract View 3 times Citations ?

Implementasi Metode SMART (Simple Multi Attribute Rating Technique) Pada Sistem Pendukung Keputusan Pemberian Kredit Pinjaman

1756-1766

Wildan Muhammad Ardana (Universitas Amikom Yogyakarta, Yogyakarta, Indonesia)
Irma Rofni Wulandari (Universitas Amikom Yogyakarta, Yogyakarta, Indonesia)
Yuli Astuti (Universitas Amikom Yogyakarta, Yogyakarta, Indonesia)
Lilis Dwi Farida (Universitas Amikom Yogyakarta, Yogyakarta, Indonesia)
Wiwi Widayani (Universitas Amikom Yogyakarta, Yogyakarta, Indonesia)

DOI: 10.30865/mib.v6i3.4333 Abstract View 11 times Citations ?

Analisis Sentimen Ulasan Hotel Bahasa Indonesia Menggunakan Support Vector Machine dan TF-IDF

1767-1774

Vincentius Westley Dimitrius Thomas (Universitas Widyatama, Bandung, Indonesia)
Fitrah Rumaisa (Universitas Widyatama, Bandung, Indonesia)

DOI: 10.30865/mib.v6i3.4218 Abstract View 3 times Citations ?

Algoritma K-Nearest Neighbors dan Synthetic Minority Oversampling Technique dalam Prediksi Pemesanan Tiket Pesawat

1775-1781

Wulan Suci (Universitas Islam Negeri Sumatera Utara, Medan, Indonesia)
Samsudin Samsudin (Universitas Islam Negeri Sumatera Utara, Medan, Indonesia)

DOI: 10.30865/mib.v6i3.4374 Abstract View 7 times Citations ?

Penerapan Metode Metode Multy Attribute Utility Theory (MAUT) dalam Pemilihan Asisten Laboratorium Komputer

1782-1788

Rima Tamara Aldisa (Universitas Nasional, Jakarta, Indonesia)
Sanwani Sanwani (Universitas Nusa Mandiri, Jakarta, Indonesia)
Deby Monalisa Simanjuntak (Universitas Budi Darma, Medan, Indonesia)
Sarpita Laia (Universitas Budi Darma, Medan, Indonesia)
Mesran Mesran (Universitas Budi Darma, Medan, Indonesia)

DOI: 10.30865/mib.v6i3.4171 Abstract View 24 times Citations ?

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1290-1296
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4019

Comparative Analysis of Multinomial Naïve Bayes and Logistic Regression Models for Prediction of SMS Spam

Pradana Ananda Raharja*, Muhammad Fajar Sidiq, Diandra Chika Fransisca

Faculty of Informatics, Informatics, Institut Teknologi Telkom Purwokerto, Banyumas, Indonesia
Email: [1,*]pradana@ittelkom-pwt.ac.id, [2]fajar@ittelkom-pwt.ac.id, [3]diandra@ittelkom-pwt.ac.id
Correspondence Author Email: pradana@ittelkom-pwt.ac.id

Abstract−This research was conducted based on a report from the United States Federal Trade Commission regarding fraud through electronic text messages via SMS that fraudsters use to manipulate potential victims. Usually, scammers spread SMS spam as an intermediary for the crime. The development of a supervised learning algorithm is applied to predict SMS spam into three categories, such as SMS spam, SMS fraud, and promotional SMS. The prediction system is dividing into several stages in the development process, including data labelling, data preprocessing, modelling, and model validation. The known accuracy based on modelling using Logistic Regression using a test size of 15% is 99%, using a test size of 20% is 99%, and using a test size of 25% is 98%. The Multinomial Naïve Bayes algorithm's accuracy with a test size of 15%, 20%, 25% is 97%. So, the SMS spam prediction approach uses the logistic regression method, which has the highest accuracy.

Keywords: Fraud; SMS Spam; Supervised Learning; Model Validation

1. INTRODUCTION

The United States Federal Trade Commission states that fraud involves sending fake text messages to trick someone into providing personal information such as passwords, account numbers, and identification numbers. Fraudsters use this information to access email or bank accounts or sell victim information to other fraudsters. Fraudsters use a variety of changing scenarios to try to get the victim's attention. Standard methods include promising gifts, gift cards, or coupons and offering low or no interest credit cards. Scammers usually send fake messages stating that they have information about the victim's account or transaction. The mode used usually says that the fraudster saw some suspicious activity on the victim's account, made a claim that there was a problem with payment information, sent fake invoices, and told the victim to contact the fraudster if the victim was going to cancel the purchase. There was even an incident where a fraudster sent a victim a fake package delivery notification[1][2]. According to the spam statistics submitted by AV-TEST, Indonesia is ranked 8th out of the world's total population in the world for global spam. The law regarding the spread of spam in Indonesia is *Undang-Undang No. 11 Tahun 2008 / Undang-Undang Informasi dan Transaksi Elektronik* (UU ITE) has not been explicitly implementing. However, sending spam can be categorized as prohibited in chapter VII article 27-34, to be precise in article 33[3][4]. Short Message Service (SMS) has developed over the decades so that it is used for business activities. SMS containing text messages is more effective than email. [5]. So that SMS is used as a tool to commit crimes and lure victims into manipulating the victim's condition [6][7].

Research conducted by Sudibyo et al. regarding the classification of spam attack attributes on email using the Decision Tree approach. Research on spam attacks with a spam dataset of 4601 records consisting of 1813 records considered spam and not spam data 278 with an initial attribute of 57 with class 1 details. One carried out three testing experiments with 30%, 50%, and 70% attribute results from unique point feature 70% better result obtained from 30% or 50% with an accuracy value of 92,469%[8]. The research conducted by Fitrian et al. aims to create an email filtering application that utilizes the naive Bayes classifier method to classify email types, including SPAM or HAM emails, and lemmatization to process words into essential words. The test results used 131 email samples, and 119 files were successfully classified correctly and while the 12 files tested got the wrong prediction value. The accuracy value obtained in this study was 90.83%[9]. Research conducted by Setiyono and Pardede investigates various data mining techniques, namely Support Vector Machine, Multinomial Naïve Bayes, and Decision Tree for automatic spam detection. Our experimental results show that the Support Vector Machine algorithm is the best of the three evaluated algorithms. Support Vector Machine reached 98.33%, while Multinomial Naïve Bayes reached 98.13% and Decision Tree with 97.10% accuracy[10]. This research was developed by evaluating the comparison of algorithms and datasets so that the aim is to compare other approaches to have a more optimum accuracy of prediction.

The development of computational methods for identifying various SMS in cyberspace requires analyse different SMS patterns[11][12]. Then make predictions against spam using processed datasets[13]. In developing a data-based SMS spam detection model, we can use techniques of machine learning. However, the prediction of SMS spam using machine learning algorithms has limitations on identifying double classification results, which means it depends on the data's characteristics[14]. Analyse several machine learning algorithms in the SMS spam detection system is to protect users from cybercrime[15]. In connection with this research, several popular machine learning classification techniques are applied, including Logistic Regression (LR) and Multinomial Naïve Bayes (MNB), to provide intelligent services in information and communication technology[16][17].

Submitted: **13/04/2022**; Accepted: **17/05/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1290-1296
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4019

The algorithm's effectiveness is tested by conducting experiments on SMS spam datasets consisting of 3 SMS categories and evaluating the algorithm's effectiveness by measuring the performance of metrics precision, recall, f1-score, and accuracy for a machine learning-based SMS spam detection model[18].

2. RESEARCH METHODOLOGY

Describe the research sequence, including research design, explain data pre-processing to process text data, make predictions using machine learning-based modelling, and model validation to determine accuracy, precision, recall, and f1-score. The explanation of the research steps is supported by references so that the explanation can be accepted scientifically. The datasets used are SMS data with various types at the data selection stage, then sorted into three data categories, including original SMS, SMS Fraud SMS, and SMS Promo. Then the pre-processing data in this study intends to process text, such as removing punctuation marks, changing to lowercase, and removing stopwords. Then the text data that has gone through the preprocessing stage is transformed into an array to be easily read by the applied algorithm. Finally, its goal is to predict text based on its category at the data mining stage. This stage aims to predict new text data not yet in the datasets. Prediction results also need to be evaluated using a confusion matrix approach to determine how accurate the method used in making predictions is. As for what needs to know that the SMS spam datasets in this study have obtained permission from previous researchers to conduct development research, using the Knowledge Discovery and Data Mining (KDD) methodology [19]. The following are the research steps carried out in extracting SMS spam text data, shown in Figure 1. The process carried out during the study consisted of the following stages.

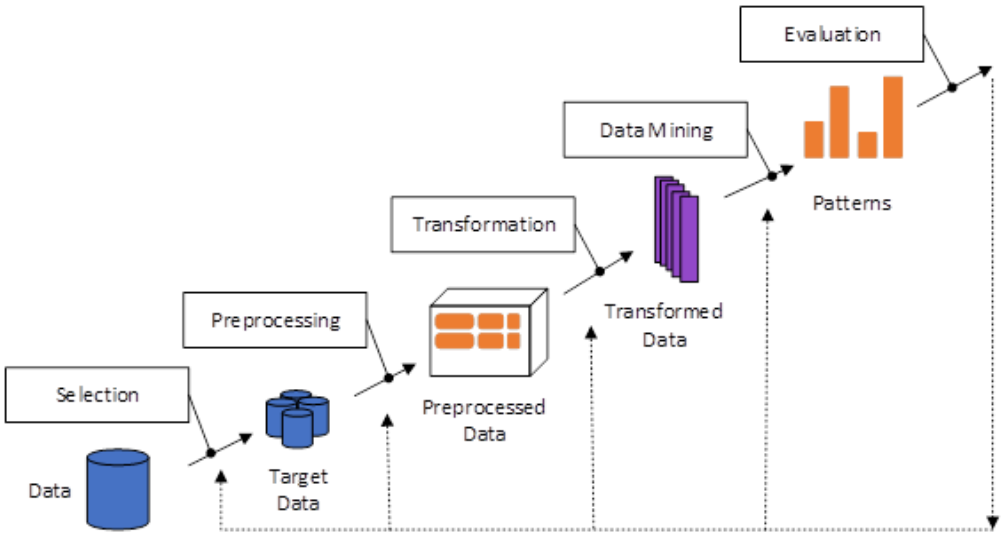


Figure 1. Data Preprocessing, Modelling, And Model Validation

2.1 Selection and Pre-processing

Selection and pre-processing are essential part of research that develops machine learning-based modelling and takes part in the analytical pipeline as our research method. The importance of applying pre-processing data in machine learning-based modelling to obtain the expected performance results[20]. The pre-processing data consisted of datasets availability, tokenization, case-folding, stop word removal, stemming, and vectorization[21][22].

a. Datasets Availability
The dataset we use in this study is SMS spam data that should make labelling by type. There are three types of SMS labels: label 0 the original SMS, label 1 is a fraud, and label 2 is SMS promotion[23]. Datasets are several datasets repositories that have information content and have relevance to research. So that data can be used to support research to be carried out[24].

b. Tokenization and Case-folding
In general, at the initial stage, the data text consists of a set of characters, and the text analysis process requires words that are available in the data set. Tokenization is simply because the text is already saved in a format that a machine can read. However, there are problems such as punctuation marks so that that punctuation marks will be removed at the tokenization stage[25]. Case-folding is briefly changing capital letters to lowercase letters to prevent ambiguity in the engine, so engine performance becomes more efficient[26].

c. Stopwords removal
One of the text processing processes in retrieving information in text or text mining or better known as stopwords removal is by deleting text from irrelevant words for indexing. There are many types of words in-

Submitted: **13/04/2022**; Accepted: **17/05/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1290-1296
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4019

text documents, such as prepositions, conjunctions, pronouns, adjectives, Etc. Some of these words may not index the document because they are not unique or never used in the search query. Therefore, this process of filtering out words is carried out—filter by providing a stoplist list. Zipf's law is sometimes used as the basis for forming non-indexable word lists, especially in the analysis of the occurrence of words[27][28].

d. *Stemming*
The stemming process is a method for extracting a word into a root word by removing all word affixes. The prefixes include prefix, suffix, and confix[29]. The application of stemming in each language has differences depending on the morphology of each language. The result of the stemming process is stem.

2.2 Transformation

Vectorization is part of data transformation, vectorization is the last stage in pre-processing data, namely changing the form of the word represented into a number[30]. The vectorization stage uses the Term Frequency - Inverse Document Frequency (TF-IDF) method to obtain each token's weight in the vector dataset. Equation (1) is a form of the TF-IDF equation carried out on each token[31].

$$w_{t,d} = tf_{t,d} \times log \frac{N}{d_{ft}} \quad (1)$$

$tf_{t,d}$: the number of occurrences of the token *t* on the document *d.*
df_t : number of documents containing tokens *t.*
N : total documents.

2.3 Data Mining

This case study uses two-approach models as a comparison, namely LR and MNB. Modelling utilizing text classification of SMS spam is using to obtain information about fraudulent SMS messages, promo SMS messages or original SMS messages[32]. Before modelling, the datasets were testing to obtain the right level of accuracy[33]. *Logistic Regression* is a supervised learning algorithm used to classify individuals based on a logistic function. Equation (2) is an equation of LR[34].

$$ln\left(\frac{p}{1-p}\right) = B_0 + B_1X \quad (2)$$

ln : natural logarithm
B_0+B_1X : the equation known as Ordinary Least Square
P : *logistic probability*

The way MNB works is to calculate the frequency of each token appearance from the document. The document sequence of occurrences of words in the document is not to account, so the document or "*bag of word*" is processed using a multinomial distribution with equation (3)[35]. Sanity check is a testing mechanism to identify valid input data after modelling[36].

$$P(c|d) = P(c)\prod_{i=1}^{n} P(w_i|c) \quad (3)$$

$P(c|d)$: class opportunity c based on the document d, n is the total number of words in the document.

$P(c) = \frac{N_c}{N}$: opportunity class c, c is class N_c is the number of class documents c, N is the number of all documents.

$P(w_i|c) = \frac{count(w_i,c)+1}{count(c)+|V|}$: the probability of the i word in class c, $count$ (w_i, c) is the number of words to x in class c, c (c) is the total number of words in class c, $|V|$ is the number of unique words in all classes.

2.4 Evaluation

The method that is generally using calculate the accuracy in machine learning in this study is the Confusion Matrix., the Confusion Matrix loads correctly predicted classification information through the classification model. The parameters used include precision, recall, f1-score, and accuracy[37].

3. RESULT AND DISCUSSION

Based Based on the results of research conducted using methods with data pre-processing stages, modelling and model validation. The research conducted by Rami and Wibisono used SMS datasets that were label as many as 1143 messages with 569 original SMS information, 335 SMS frauds, and 239 SMS promos shown in Figure 2. The modelling applied in this study uses two supervised learning methods, namely, LR and MNB.

Submitted: **13/04/2022**; Accepted: **17/05/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1290-1296
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4019

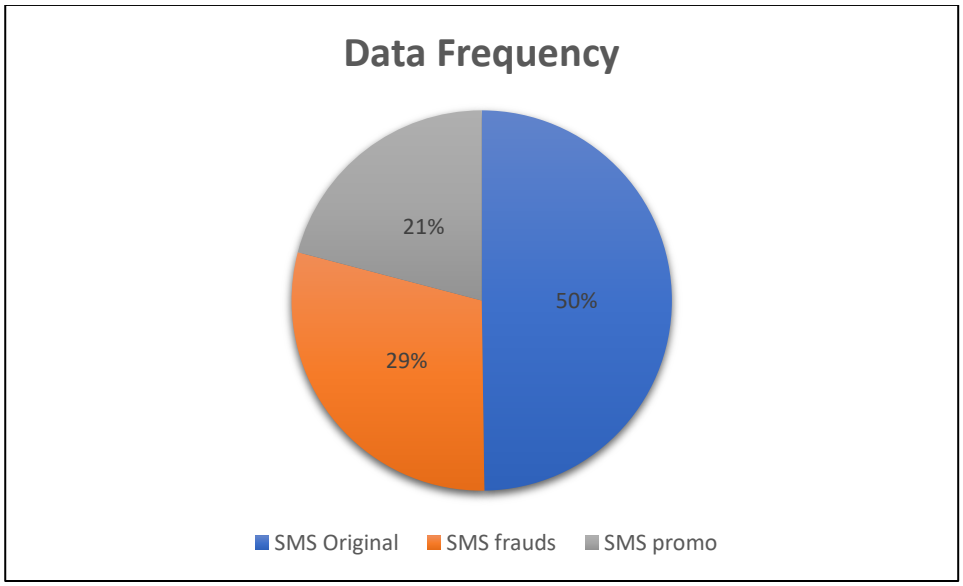


Figure 2. Datasets SMS

3.1 Selection, Pre-processing and Transformation

The data pre-processing stage consists of tokenization, case-folding, stopwords removal, stemming, and vectorization using libraries available in the Python programming language, which shown in Figure 3. Figure 4 is the output of data pre-processing which has been in the form of vectors.

```
import nltk
nltk.download('stopwords')
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from string import punctuation
sw_indo = stopwords.words("indonesian") +
list(punctuation)
```

Figure 3. Library for data pre-processing

```
array ([[0, 0, 0, ..., 0, 0, 0],
        [0, 0, 0, ..., 0, 0, 0],
        [0, 0, 0, ..., 0, 0, 0],
        ...,
        [0, 0, 0, ..., 0, 0, 0],
        [0, 0, 0, ..., 0, 0, 0],
      [0, 0, 0, ..., 0, 0, 0]])
```

Figure 4. The output of data transformation

3.2 Data mining

Prediction of modelling variation to predict three SMS text classifications using LR and MNB supported by the scikit-learn library by testing dataset sizes of 15%, 20%, and 25% of the total data and accompanied by the results of checking the accuracy of prediction algorithms, which following in Table 1.

Table 1. Results of Prediction using Logistic Regression and Multinomial Naïve Bayes

Phone Number	Sample SMS	Predictions	Method and Weight Percentage of Tested Datasets					
			Logistic Regression			Multinominal Naïve Bayes		
			15%	20%	25%	15%	20%	25%
6282299209***	Maaf Mengganggu Waktunya KAMI KOPERASI Menawarkan PNJMN-ONLINE 5jt Sampai 5OOjt Bunga 4% Pertahun Cepat & Mudah INFO WhatsApp: +6285298436***	Fraud	70,40%	56,59%	52,04%	98,78%	99,14%	99,48%
6285238123***	YTH BPK/IBU KMI MELAYANI PENGAJUAN RUPIAH CEPAT DGN PROSES CEPAT TAMPA ANGGUNAN	Fraud	99,20%	94,75%	82,12%	99,99%	99,99%	99,99%

Submitted: **13/04/2022**; Accepted: **17/05/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1290-1296
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4019

Phone Number	Sample SMS	Predictions	Method and Weight Percentage of Tested Datasets					
			Logistic Regression			Multinominal Naïve Bayes		
			15%	20%	25%	15%	20%	25%
CFC	MINIMAL 5jt-500jt INFO LENKAP HUB KMI DI WA;0823-9805-0***	Fraud	99,51%	96,22%	87,50%	100%	100%	100%
	Bpk/Ibu Mengenai Rekening Anda Terpilih Sebagai Pemenang Cek 35jt Dri BNI U/Info klik www.promobni46.tk Kode Cek 03299757							
	Hub.085288991*** DISKON 40%. 2 Ayam + 2 Chicken Strips + 2 Nasi hanya 39 RIBU NETT. Tukar SMS di CFC STASIUN PURWOKERTO hingga 14 Des. SKB. Promo *606#	Promo	96,50%	80,39%	65,30%	99,99%	99,99%	99,99%
Tokopedia	Bayar PBB gak pake antri! Cashback s.d Rp10.000 dengan kode promo: GEBYARPBB Hanya di tsel.me/pbbtokped	Promo	65,65%	51,52%	45,05%	99,99%	99,99%	99,99%
Starbucks	BELI 1 GRATIS 1. HANYA HARI INI. Semua Minuman! Tall Size! Tukarkan SMS hari ini di Starbucks terdekat (exc.Airport). S&K Berlaku. Promo*606#	Promo	99,91%	97,48%	96,79%	99,99%	99,99%	99,99%
…	…	…	…	…	…	…	…	…
085229991***	bntr lagi pulang	Original SMS	97,81%	91,01%	46,74%	99,06%	99,01%	98,82%

3.3 Evaluation

Then the accuracy performance test results by dividing the datasets sorted from lowest to highest accuracy, namely the MNB method, with datasets of 75%, 80%, and 85% having an accuracy rate of 97%. While the LR algorithm has better results, namely on datasets, 75% have an accuracy of 98%, 80% have an accuracy of 99%, and 85% have an accuracy of 99%, as shown in Figure 5.

Table 2. Evaluation of Classification Performance with Datasets Ratio

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
LR 15%	99	98	99	99
LR 20%	99	99	99	99
LR 25%	98	98	98	98
MNB 15%	97	97	97	97
MNB 20%	97	97	97	97
MNB 25%	97	97	97	97

Submitted: **13/04/2022**; Accepted: **17/05/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1290-1296
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4019

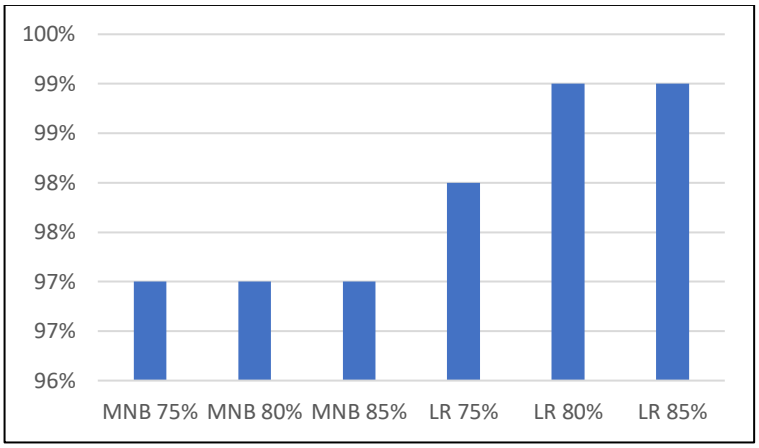


Figure 5. *Accuracy Comparison of Classification Performance*

4. CONCLUSION

Based on results of research that has been done with validation using confusion matrix, the conclusion of the LR algorithm with a test size of 15% has an accuracy of 99%, a test size of 20% has an accuracy of 99%, and a test size of 25% has an accuracy of 98%. The MNB algorithm with a test size of 15%, 20%, 25% has the same accuracy, namely 97%. With the information obtained from this study, the LR algorithm has the best accuracy in making predictions.

REFERENCES

[1] United State of America Federal Trade Commision, "How to Recognize and Report Spam Text Messages," *Consumer Information*, 2020. https://www.consumer.ftc.gov/articles/how-recognize-and-report-spam-text-messages (accessed Dec. 12, 2020).
[2] O. S. Yee, S. Sagadevan, and N. H. A. H. Malim, "Credit Card Fraud Detection Using Machine Learning As Data Mining Technique," *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 10, no. 1–4, pp. 23–27, 2018.
[3] Y. Vernanda, S. Hansun, and M. B. Kristanda, "Indonesian language email spam detection using N-gram and Naïve Bayes algorithm," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 5, pp. 2012–2019, 2020, doi: 10.11591/eei.v9i5.2444.
[4] M. Rifauddin and A. N. Halida, "Waspada Cybercrime dan Informasi Hoax Pada Media Sosial Facebook," *Khizanah al-Hikmah : Jurnal Ilmu Perpustakaan, Informasi, dan Kearsipan*, vol. 6, no. 2, pp. 98–111, 2018, doi: 10.24252/kah.v6i2a2.
[5] P. K. Roy, J. P. Singh, and S. Banerjee, "Deep learning to filter SMS Spam," *Future Generation Computer Systems*, vol. 102, pp. 524–533, 2020, doi: 10.1016/j.future.2019.09.001.
[6] I. Rahmawati, "Analisis Manajemen Resiko Ancaman Kejahatan Siber (Cyber Crime) Dalam Peningkatan Cyber Defense," *Jurnal Pertahanan & Bela Negara*, vol. 7, no. 2, pp. 51–66, 2017, doi: 10.33172/jpbh.v7i2.193.
[7] R. C. Perkins, C. J. Howell, C. E. Dodge, G. W. Burruss, and D. Maimon, "Malicious Spam Distribution: A Routine Activities Approach," *Deviant Behavior*, vol. 00, no. 00, pp. 1–17, 2020, doi: 10.1080/01639625.2020.1794269.
[8] A. Sudibyo, T. Asra, and B. Rifai, "Klasifikasi Seleksi Atribut Pada Serangan Spam Menggunakan Metode Algoritma Decision Tree," *Jurnal PILAR Nusa Mandiri*, vol. 14, no. 2, pp. 145–150, 2018, [Online]. Available: http://nusamandiri.ac.id/aji.aby@nusamandiri.ac.idhttp://bsi.ac.idhttp://nusamandiri.ac.id/
[9] H. P. Fitrian, I. Ruslianto, and R. Hidayat, "Implementasi Metode Naive Bayes Classifier Untuk Aplikasi Filtering Email Spam Dengan Lemmatization Berbasis Web," *Jurnal Coding, Sistem Komputer Untan*, vol. 06, no. 02, pp. 13–24, 2018.
[10] A. Setiyono and H. F Pardede, "Klasifikasi Sms Spam Menggunakan Support Vector Machine," *Jurnal Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 275–280, Sep. 2019, doi: 10.33480/pilar.v15i2.693.
[11] D. Kawade and K. Oza, "Content-Based SMS Spam Filtering Using Machine Learning Technique," *International Journal of Computer Engineering and Applications*, vol. 13, no. 4, 2018.
[12] M. Bassiouni, M. Ali, and E. A. El-Dahshan, "Ham and Spam E-Mails Classification Using Machine Learning Techniques," *Journal of Applied Security Research*, vol. 13, no. 3, pp. 315–331, 2018, doi: 10.1080/19361610.2018.1463136.
[13] A. K. Jain, S. K. Yadav, and N. Choudhary, "A novel Approach to Detect Spam and Smishing SMS using Machine Learning Techniques," *International Journal of E-Services and Mobile Applications*, vol. 12, no. 1, pp. 21–38, 2020, doi: 10.4018/IJESMA.2020010102.
[14] N. K. Nagwani and A. Sharaff, "SMS Spam Filtering and Thread Identification using Bi-Level Text Classification and Clustering Techniques," *Journal of Information Science*, vol. 43, no. 1, pp. 1–13, 2017, doi: 10.1177/0165551515616310.
[15] A. Ghourabi, M. A. Mahmood, and Q. M. Alzubi, "A hybrid CNN-LSTM model for SMS spam detection in arabic and english messages," *Future Internet*, vol. 12, no. 9, pp. 1–16, 2020, doi: 10.3390/FI12090156.
[16] M. Manap, M. H. Jopri, A. R. Abdullah, R. Karim, M. R. Yusoff, and A. H. Azahar, "A verification of periodogram technique for harmonic source diagnostic analytic by using logistic regression," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 17, no. 1, pp. 497–507, 2019, doi: 10.12928/TELKOMNIKA.v17i1.10390.

Submitted: **13/04/2022**; Accepted: **17/05/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1290-1296
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4019

[17] N. Shiri Harzevili and S. H. Alizadeh, "Mixture of Latent Multinomial Naïve Bayes Classifier," *Applied Soft Computing Journal*, vol. 69, pp. 516–527, 2018, doi: 10.1016/j.asoc.2018.04.020.
[18] J. Feldman, A. Thomas-Bachli, J. Forsyth, Z. H. Patel, and K. Khan, "Development of a Global Infectious Disease Activity Database using Natural Language Processing, Machine Learning, and Human Expertise," *Journal of the American Medical Informatics Association*, vol. 26, no. 11, pp. 1355–1359, 2019, doi: 10.1093/jamia/ocz112.
[19] H. M. Safhi, B. Frikh, and B. Ouhbi, "Assessing reliability of Big Data Knowledge Discovery process," *Procedia Computer Science*, vol. 148, pp. 30–36, 2019, doi: 10.1016/j.procs.2019.01.005.
[20] X. Zheng, M. Wang, and J. Ordieres-Meré, "Comparison of Data Preprocessing Approaches for Applying Deep Learning to Human Activity Recognition in the Context of Industry 4.0," *Sensors (Switzerland)*, vol. 18, no. 7, 2018, doi: 10.3390/s18072146.
[21] S. Khomsah and Agus Sasmito Aribowo, "Model Text-Preprocessing Komentar Youtube Dalam Bahasa Indonesia," *Rekayasa Sistem dan Teknologi Informasi, RESTI*, vol. 4, no. 10, pp. 648–654, 2020.
[22] W. T. H. Putri, M. S. Prastio, R. Hendrowati, Y. Sari, and H. T. Y. Achsan, "Content-based Filtering Model for Recommendation of Indonesian Legal Article Study Case of Klinik Hukumonline," in *2019 International Workshop on Big Data and Information Security, IWBIS 2019*, 2019, pp. 9–14. doi: 10.1109/IWBIS.2019.8935726.
[23] F. Rahmi and W. Yudi, "Aplikasi SMS Spam Filtering pada Android menggunakan Naïve Bayes," Universitas Pendidikan Indonesia, 2017.
[24] S. R. Kunze and S. Auer, "Dataset retrieval," in *Proceedings - 2013 IEEE 7th International Conference on Semantic Computing, ICSC 2013*, 2013, pp. 1–8. doi: 10.1109/ICSC.2013.12.
[25] S. Vijayarani and J. Rajaraman, "Text Mining: open Source Tokenization Tools – An Analysis," *Advanced Computational Intelligence: An International Journal (ACII)*, vol. 3, no. 1, pp. 37–47, 2016, doi: 10.5121/acii.2016.3104.
[26] C. C. Aggarwal, *Machine Learning for Text*. Yorktown Heights: Springer, 2018. doi: 10.1007/978-3-319-73531-3_10.
[27] F. Rahutomo and A. R. T. H. Ririd, "Evaluasi Daftar Stopword Bahasa Indonesia," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 6, no. 1, pp. 41–47, 2019, doi: 10.25126/jtiik.2019611226.
[28] A. F. Hidayatullah, "Pengaruh Stopword Terhadap Performa Klasifikasi Tweet Berbahasa Indonesia," *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 1, no. 1, pp. 1–4, 2016.
[29] A. B. Arifa, G. F. Fitriana, and A. R. Hasan, "Temu Kembali Informasi pada Soal Ujian dengan Rencana Pembelajaran Menggunakan Vector Space Model," *Jurnal Resti*, vol. 5, no. 1, pp. 8–12, 2021.
[30] L. A. Wirasakti, R. Permadi, A. D. Hartanto, and H. Hartatik, "Pembuatan Kata Kunci Otomatis Dalam Artikel Dengan Pemodelan Topik," *Jurnal Media Informatika Budidarma*, vol. 4, no. 1, p. 27, 2020, doi: 10.30865/mib.v4i1.1707.
[31] N. Abdulloh and A. F. Hidayatullah, "Deteksi Cyberbullying pada Cuitan Media Sosial Twitter," *Automata*, vol. Vol 1, no. 1, pp. 1–5, 2019.
[32] L. Mutawalli, M. T. A. Zaen, and W. Bagye, "Klasifikasi Teks Sosial Media Twitter Menggunakan Support Vector Machine (Studi Kasus Penusukan Wiranto)," *Jurnal Informatika dan Rekayasa Elektronik*, vol. 2, no. 2, pp. 43–51, 2019, doi: 10.36595/jire.v2i2.117.
[33] A. Santoso and G. Ariyanto, "Implementasi Deep Learning Berbasis Keras untuk Pengenalan Wajah," *Emitor*, vol. 18, no. 01, pp. 15–21, 2018, doi: 10.23917/emitor.v18i01.6235.
[34] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification," *Augmented Human Research*, vol. 5, no. 1, pp. 1–16, 2020, doi: 10.1007/s41133-020-00032-0.
[35] S. Fanissa, M. A. Fauzi, and S. Adinugroho, "Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking | Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 8, pp. 2766–2770, 2018.
[36] H. Lu, H. Xu, N. Liu, Y. Zhou, and X. Wang, "Data sanity check for deep learning systems via learnt assertions," in *ASE 2019*, 2019, pp. 1–3.
[37] E. Indrayuni, "Klasifikasi Text Mining Review Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes," *Jurnal Khatulistiwa Informatika*, vol. 7, no. 1, pp. 29–36, 2019, doi: 10.31294/jki.v7i1.1.

Submitted: **13/04/2022**; Accepted: **17/05/2022**; Published: **25/07/2022**

PAPER NAME

4019-12539-2-PB.pdf

AUTHOR

Pradana Raharja

WORD COUNT

4393 Words

CHARACTER COUNT

23558 Characters

PAGE COUNT

7 Pages

FILE SIZE

513.9KB

SUBMISSION DATE

Jul 23, 2022 9:19 PM GMT+7

REPORT DATE

Jul 23, 2022 9:20 PM GMT+7

6% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 4% Internet database
- 4% Publications database
- Crossref database
- Crossref Posted Content database
- 2% Submitted Works database

Excluded from Similarity Report

- Bibliographic material
- Quoted material
- Cited material
- Small Matches (Less then 8 words)
- Manually excluded text blocks

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1290-1296
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4019

Comparative Analysis of Multinomial Naïve Bayes and Logistic Regression Models for Prediction of SMS Spam

Pradana Ananda Raharja*, Muhammad Fajar Sidiq, Diandra Chika Fransisca

Faculty of Informatics, Informatics, Institut Teknologi Telkom Purwokerto, Banyumas, Indonesia
Email: [1,*]pradana@ittelkom-pwt.ac.id, [2]fajar@ittelkom-pwt.ac.id, [3]diandra@ittelkom-pwt.ac.id
Correspondence Author Email: pradana@ittelkom-pwt.ac.id

Abstract−This research was conducted based on a report from the United States Federal Trade Commission regarding fraud through electronic text messages via SMS that fraudsters use to manipulate potential victims. Usually, scammers spread SMS spam as an intermediary for the crime. The development of a supervised learning algorithm is applied to predict SMS spam into three categories, such as SMS spam, SMS fraud, and promotional SMS. The prediction system is dividing into several stages in the development process, including data labelling, data preprocessing, modelling, and model validation. The known accuracy based on modelling using Logistic Regression using a test size of 15% is 99%, using a test size of 20% is 99%, and using a test size of 25% is 98%. The Multinomial Naïve Bayes algorithm's accuracy with a test size of 15%, 20%, 25% is 97%. So, the SMS spam prediction approach uses the logistic regression method, which has the highest accuracy.

Keywords: Fraud; SMS Spam; Supervised Learning; Model Validation

1. INTRODUCTION

The United States Federal Trade Commission states that fraud involves sending fake text messages to trick someone into providing personal information such as passwords, account numbers, and identification numbers. Fraudsters use this information to access email or bank accounts or sell victim information to other fraudsters. Fraudsters use a variety of changing scenarios to try to get the victim's attention. Standard methods include promising gifts, gift cards, or coupons and offering low or no interest credit cards. Scammers usually send fake messages stating that they have information about the victim's account or transaction. The mode used usually says that the fraudster saw some suspicious activity on the victim's account, made a claim that there was a problem with payment information, sent fake invoices, and told the victim to contact the fraudster if the victim was going to cancel the purchase. There was even an incident where a fraudster sent a victim a fake package delivery notification[1][2]. According to the spam statistics submitted by AV-TEST, Indonesia is ranked 8th out of the world's total population in the world for global spam. The law regarding the spread of spam in Indonesia is *Undang-Undang No. 11 Tahun 2008 / Undang-Undang Informasi dan Transaksi Elektronik* (UU ITE) has not been explicitly implementing. However, sending spam can be categorized as prohibited in chapter VII article 27-34, to be precise in article 33[3][4]. Short Message Service (SMS) has developed over the decades so that it is used for business activities. SMS containing text messages is more effective than email. [5]. So that SMS is used as a tool to commit crimes and lure victims into manipulating the victim's condition [6][7].

Research conducted by Sudibyo et al. regarding the classification of spam attack attributes on email using the Decision Tree approach. Research on spam attacks with a spam dataset of 4601 records consisting of 2813 records considered spam and not spam data 278 with an initial attribute of 57 with class 1 details. One carried out three testing experiments with 30%, 50%, and 70% attribute results from unique point feature 70% better result obtained from 30% or 50% with an accuracy value of 92,469%[8]. The research conducted by Fitrian et al. aims to create an email filtering application that utilizes the naive Bayes classifier method to classify email types, including SPAM or HAM emails, and lemmatization to process words into essential words. The test results used 131 email samples, and 119 files were successfully classified correctly and while the 12 files tested got the wrong prediction value. The accuracy value obtained in this study was 90.83%[9]. Research conducted by Setiyono and Pardede investigates various data mining techniques, namely Support Vector Machine, Multinomial Naïve Bayes, and Decision Tree for automatic spam detection. Our experimental results show that the Support Vector Machine algorithm is the best of the three evaluated algorithms. Support Vector Machine reached 98.33%, while Multinomial Naïve Bayes reached 98.13% and Decision Tree with 97.10% accuracy[10]. This research was developed by evaluating the comparison of algorithms and datasets so that the aim is to compare other approaches to have a more optimum accuracy of prediction.

The development of computational methods for identifying various SMS in cyberspace requires analyse different SMS patterns[11][12]. Then make predictions against spam using processed datasets[13]. In developing a data-based SMS spam detection model, we can use techniques of machine learning. However, the prediction of SMS spam using machine learning algorithms has limitations on identifying double classification results, which means it depends on the data's characteristics[14]. Analyse several machine learning algorithms in the SMS spam detection system is to protect users from cybercrime[15]. In connection with this research, several popular machine learning classification techniques are applied, including Logistic Regression (LR) and Multinomial Naïve Bayes (MNB), to provide intelligent services in information and communication technology[16][17].

Submitted: **13/04/2022**; Accepted: **17/05/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1290-1296
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4019

The algorithm's effectiveness is tested by conducting experiments on SMS spam datasets consisting of 3 SMS categories and evaluating the algorithm's effectiveness by measuring the performance of metrics precision, recall, f1-score, and accuracy for a machine learning-based SMS spam detection model[18].

2. RESEARCH METHODOLOGY

Describe the research sequence, including research design, explain data pre-processing to process text data, make predictions using machine learning-based modelling, and model validation to determine accuracy, precision, recall, and f1-score. The explanation of the research steps is supported by references so that the explanation can be accepted scientifically. The datasets used are SMS data with various types at the data selection stage, then sorted into three data categories, including original SMS, SMS Fraud SMS, and SMS Promo. Then the pre-processing data in this study intends to process text, such as removing punctuation marks, changing to lowercase, and removing stopwords. Then the text data that has gone through the preprocessing stage is transformed into an array to be easily read by the applied algorithm. Finally, its goal is to predict text based on its category at the data mining stage. This stage aims to predict new text data not yet in the datasets. Prediction results also need to be evaluated using a confusion matrix approach to determine how accurate the method used in making predictions is. As for what needs to know that the SMS spam datasets in this study have obtained permission from previous researchers to conduct development research, using the Knowledge Discovery and Data Mining (KDD) methodology [19]. The following are the research steps carried out in extracting SMS spam text data, shown in Figure 1. The process carried out during the study consisted of the following stages.

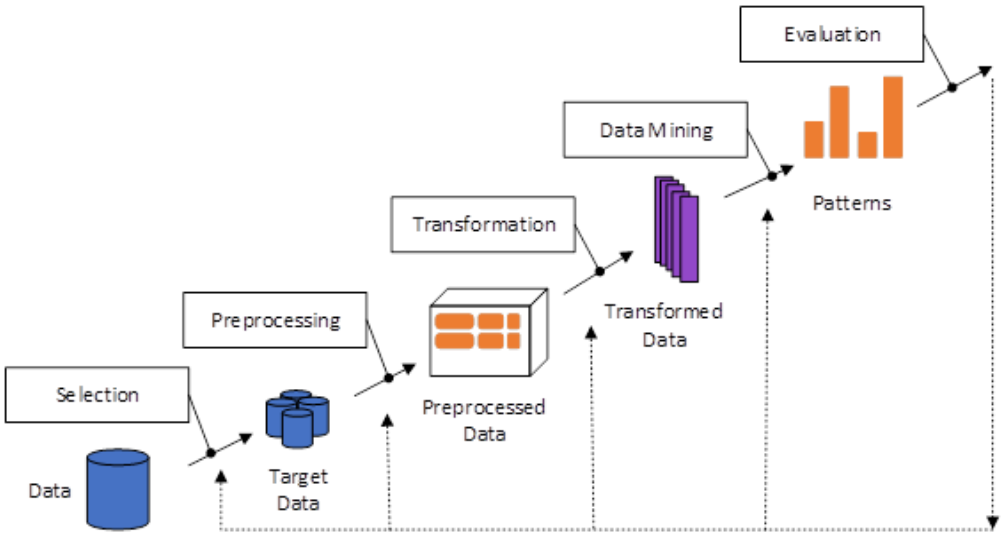


Figure 1. Data Preprocessing, Modelling, And Model Validation

2.1 Selection and Pre-processing

Selection and pre-processing are essential part of research that develops machine learning-based modelling and takes part in the analytical pipeline as our research method. The importance of applying pre-processing data in machine learning-based modelling to obtain the expected performance results[20]. The pre-processing data consisted of datasets availability, tokenization, case-folding, stop word removal, stemming, and vectorization[21][22].

a. Datasets Availability
The dataset we use in this study is SMS spam data that should make labelling by type. There are three types of SMS labels: label 0 the original SMS, label 1 is a fraud, and label 2 is SMS promotion[23]. Datasets are several datasets repositories that have information content and have relevance to research. So that data can be used to support research to be carried out[24].

b. Tokenization and Case-folding
In general, at the initial stage, the data text consists of a set of characters, and the text analysis process requires words that are available in the data set. Tokenization is simply because the text is already saved in a format that a machine can read. However, there are problems such as punctuation marks so that that punctuation marks will be removed at the tokenization stage[25]. Case-folding is briefly changing capital letters to lowercase letters to prevent ambiguity in the engine, so engine performance becomes more efficient[26].

c. Stopwords removal
One of the text processing processes in retrieving information in text or text mining or better known as stopwords removal is by deleting text from irrelevant words for indexing. There are many types of words in-

Submitted: 13/04/2022; Accepted: 17/05/2022; Published: 25/07/2022

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1290-1296
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4019

text documents, such as prepositions, conjunctions, pronouns, adjectives, Etc. Some of these words may not index the document because they are not unique or never used in the search query. Therefore, this process of filtering out words is carried out—filter by providing a stoplist list. Zipf's law is sometimes used as the basis for forming non-indexable word lists, especially in the analysis of the occurrence of words[27][28].

d. *Stemming*
The stemming process is a method for extracting a word into a root word by removing all word affixes. The prefixes include prefix, suffix, and confix[29]. The application of stemming in each language has differences depending on the morphology of each language. The result of the stemming process is stem.

2.2 Transformation

Vectorization is part of data transformation, vectorization is the last stage in pre-processing data, namely changing the form of the word represented into a number[30]. The vectorization stage uses the Term Frequency - Inverse Document Frequency (TF-IDF) method to obtain each token's weight in the vector dataset. Equation (1) is a form of the TF-IDF equation carried out on each token[31].

$$w_{t,d} = tf_{t,d} \times log \frac{N}{d_{ft}} \quad (1)$$

$tf_{t,d}$: the number of occurrences of the token *t* on the document *d.*
df_t : number of documents containing tokens *t.*
N : total documents.

2.3 Data Mining

This case study uses two-approach models as a comparison, namely LR and MNB. Modelling utilizing text classification of SMS spam is using to obtain information about fraudulent SMS messages, promo SMS messages or original SMS messages[32]. Before modelling, the datasets were testing to obtain the right level of accuracy[33]. *Logistic Regression* is a supervised learning algorithm used to classify individuals based on a logistic function. Equation (2) is an equation of LR[34].

$$ln\left(\frac{p}{1-p}\right) = B_0 + B_1X \quad (2)$$

ln : natural logarithm
B_0+B_1X : the equation known as Ordinary Least Square
P : *logistic probability*

The way MNB works is to calculate the frequency of each token appearance from the document. The document sequence of occurrences of words in the document is not to account, so the document or "*bag of word*" is processed using a multinomial distribution with equation (3)[35]. Sanity check is a testing mechanism to identify valid input data after modelling[36].

$$P(c|d) = P(c)\prod_{i=1}^{n} P(w_i|c) \quad (3)$$

$P(c|d)$: class opportunity c based on the document d, n is the total number of words in the document.

$P(c) = \frac{N_c}{N}$: opportunity class c, c is class N_c is the number of class documents c, N is the number of all documents.

$P(w_i|c) = \frac{count(w_i,c)+1}{count(c)+|V|}$: the probability of the i word in class c, $count$ (w_i, c) is the number of words to x in class c, c (c) is the total number of words in class c, $|V|$ is the number of unique words in all classes.

2.4 Evaluation

The method that is generally using calculate the accuracy in machine learning in this study is the Confusion Matrix., the Confusion Matrix loads correctly predicted classification information through the classification model. The parameters used include precision, recall, f1-score, and accuracy[37].

3. RESULT AND DISCUSSION

Based Based on the results of research conducted using methods with data pre-processing stages, modelling and model validation. The research conducted by Rami and Wibisono used SMS datasets that were label as many as 1143 messages with 569 original SMS information, 335 SMS frauds, and 239 SMS promos shown in Figure 2. The modelling applied in this study uses two supervised learning methods, namely, LR and MNB.

Submitted: **13/04/2022**; Accepted: **17/05/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1290-1296
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4019

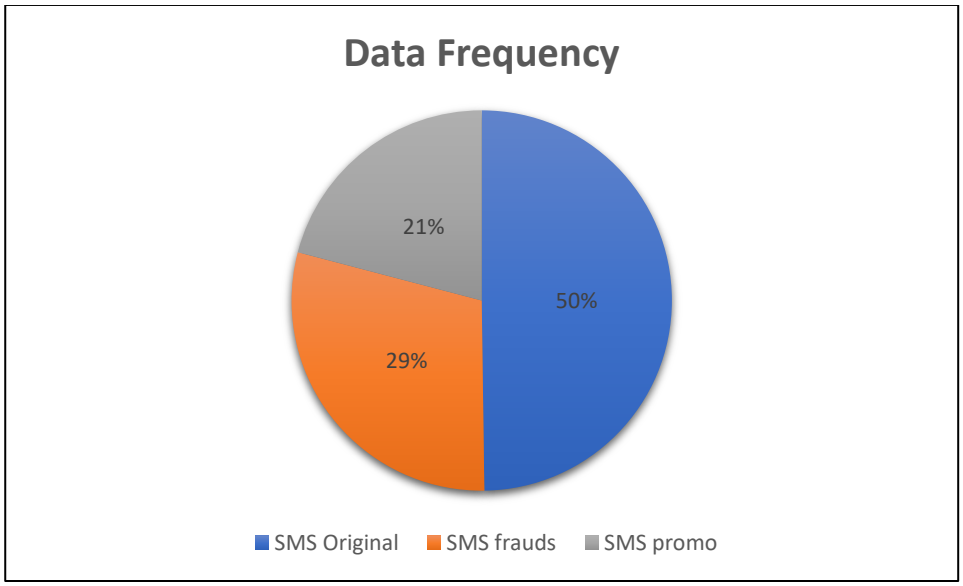


Figure 2. Datasets SMS

3.1 Selection, Pre-processing and Transformation

The data pre-processing stage consists of tokenization, case-folding, stopwords removal, stemming, and vectorization using libraries available in the Python programming language, which shown in Figure 3. Figure 4 is the output of data pre-processing which has been in the form of vectors.

```
import nltk
nltk.download('stopwords')
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from string import punctuation
sw_indo = stopwords.words("indonesian") +
list(punctuation)
```

Figure 3. Library for data pre-processing

```
array ([[0, 0, 0, ..., 0, 0, 0],
        [0, 0, 0, ..., 0, 0, 0],
        [0, 0, 0, ..., 0, 0, 0],
        ...,
        [0, 0, 0, ..., 0, 0, 0],
        [0, 0, 0, ..., 0, 0, 0],
      [0, 0, 0, ..., 0, 0, 0]])
```

Figure 4. The output of data transformation

3.2 Data mining

Prediction of modelling variation to predict three SMS text classifications using LR and MNB supported by the scikit-learn library by testing dataset sizes of 15%, 20%, and 25% of the total data and accompanied by the results of checking the accuracy of prediction algorithms, which following in Table 1.

Table 1. Results of Prediction using Logistic Regression and Multinomial Naïve Bayes

Phone Number	Sample SMS	Predictions	Method and Weight Percentage of Tested Datasets					
			Logistic Regression			Multinominal Naïve Bayes		
			15%	20%	25%	15%	20%	25%
6282299209***	Maaf Mengganggu Waktunya KAMI KOPERASI Menawarkan PNJMN-ONLINE 5jt Sampai 5OOjt Bunga 4% Pertahun Cepat & Mudah INFO WhatsApp: +6285298436***	Fraud	70,40%	56,59%	52,04%	98,78%	99,14%	99,48%
6285238123***	YTH BPK/IBU KMI MELAYANI PENGAJUAN RUPIAH CEPAT DGN PROSES CEPAT TAMPA ANGGUNAN	Fraud	99,20%	94,75%	82,12%	99,99%	99,99%	99,99%

Submitted: **13/04/2022**; Accepted: **17/05/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1290-1296
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4019

Phone Number	Sample SMS	Predictions	Method and Weight Percentage of Tested Datasets					
			Logistic Regression			Multinominal Naïve Bayes		
			15%	20%	25%	15%	20%	25%
CFC	MINIMAL 5jt-500jt INFO LENKAP HUB KMI DI WA;0823-9805-0***	Fraud	99,51%	96,22%	87,50%	100%	100%	100%
	Bpk/Ibu Mengenai Rekening Anda Terpilih Sebagai Pemenang Cek 35jt Dri BNI U/Info klik www.promobni46.tk Kode Cek 03299757							
	Hub.085288991*** DISKON 40%. 2 Ayam + 2 Chicken Strips + 2 Nasi hanya 39 RIBU NETT. Tukar SMS di CFC STASIUN PURWOKERTO hingga 14 Des. SKB. Promo *606#	Promo	96,50%	80,39%	65,30%	99,99%	99,99%	99,99%
Tokopedia	Bayar PBB gak pake antri! Cashback s.d Rp10.000 dengan kode promo: GEBYARPBB Hanya di tsel.me/pbbtokped	Promo	65,65%	51,52%	45,05%	99,99%	99,99%	99,99%
Starbucks	BELI 1 GRATIS 1. HANYA HARI INI. Semua Minuman! Tall Size! Tukarkan SMS hari ini di Starbucks terdekat (exc.Airport). S&K Berlaku. Promo*606#	Promo	99,91%	97,48%	96,79%	99,99%	99,99%	99,99%
…	…	…	…	…	…	…	…	…
085229991***	bntr lagi pulang	Original SMS	97,81%	91,01%	46,74%	99,06%	99,01%	98,82%

3.3 Evaluation

Then the accuracy performance test results by dividing the datasets sorted from lowest to highest accuracy, namely the MNB method, with datasets of 75%, 80%, and 85% having an accuracy rate of 97%. While the LR algorithm has better results, namely on datasets, 75% have an accuracy of 98%, 80% have an accuracy of 99%, and 85% have an accuracy of 99%, as shown in Figure 5.

Table 2. Evaluation of Classification Performance with Datasets Ratio

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
LR 15%	99	98	99	99
LR 20%	99	99	99	99
LR 25%	98	98	98	98
MNB 15%	97	97	97	97
MNB 20%	97	97	97	97
MNB 25%	97	97	97	97

Submitted: **13/04/2022**; Accepted: **17/05/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1290-1296
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4019

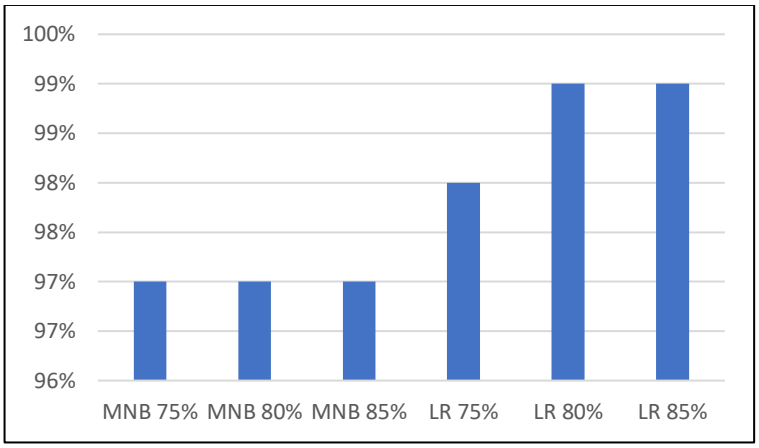


Figure 5. *Accuracy Comparison of Classification Performance*

4. CONCLUSION

Based on results of research that has been done with validation using confusion matrix, the conclusion of the LR algorithm with a test size of 15% has an accuracy of 99%, a test size of 20% has an accuracy of 99%, and a test size of 25% has an accuracy of 98%. The MNB algorithm with a test size of 15%, 20%, 25% has the same accuracy, namely 97%. With the information obtained from this study, the LR algorithm has the best accuracy in making predictions.

REFERENCES

[1] United State of America Federal Trade Commision, "How to Recognize and Report Spam Text Messages," *Consumer Information*, 2020. https://www.consumer.ftc.gov/articles/how-recognize-and-report-spam-text-messages (accessed Dec. 12, 2020).
[2] O. S. Yee, S. Sagadevan, and N. H. A. H. Malim, "Credit Card Fraud Detection Using Machine Learning As Data Mining Technique," *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 10, no. 1–4, pp. 23–27, 2018.
[3] Y. Vernanda, S. Hansun, and M. B. Kristanda, "Indonesian language email spam detection using N-gram and Naïve Bayes algorithm," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 5, pp. 2012–2019, 2020, doi: 10.11591/eei.v9i5.2444.
[4] M. Rifauddin and A. N. Halida, "Waspada Cybercrime dan Informasi Hoax Pada Media Sosial Facebook," *Khizanah al-Hikmah : Jurnal Ilmu Perpustakaan, Informasi, dan Kearsipan*, vol. 6, no. 2, pp. 98–111, 2018, doi: 10.24252/kah.v6i2a2.
[5] P. K. Roy, J. P. Singh, and S. Banerjee, "Deep learning to filter SMS Spam," *Future Generation Computer Systems*, vol. 102, pp. 524–533, 2020, doi: 10.1016/j.future.2019.09.001.
[6] I. Rahmawati, "Analisis Manajemen Resiko Ancaman Kejahatan Siber (Cyber Crime) Dalam Peningkatan Cyber Defense," *Jurnal Pertahanan & Bela Negara*, vol. 7, no. 2, pp. 51–66, 2017, doi: 10.33172/jpbh.v7i2.193.
[7] R. C. Perkins, C. J. Howell, C. E. Dodge, G. W. Burruss, and D. Maimon, "Malicious Spam Distribution: A Routine Activities Approach," *Deviant Behavior*, vol. 00, no. 00, pp. 1–17, 2020, doi: 10.1080/01639625.2020.1794269.
[8] A. Sudibyo, T. Asra, and B. Rifai, "Klasifikasi Seleksi Atribut Pada Serangan Spam Menggunakan Metode Algoritma Decision Tree," *Jurnal PILAR Nusa Mandiri*, vol. 14, no. 2, pp. 145–150, 2018, [Online]. Available: http://nusamandiri.ac.id/aji.aby@nusamandiri.ac.idhttp://bsi.ac.idhttp://nusamandiri.ac.id/
[9] H. P. Fitrian, I. Ruslianto, and R. Hidayat, "Implementasi Metode Naive Bayes Classifier Untuk Aplikasi Filtering Email Spam Dengan Lemmatization Berbasis Web," *Jurnal Coding, Sistem Komputer Untan*, vol. 06, no. 02, pp. 13–24, 2018.
[10] A. Setiyono and H. F Pardede, "Klasifikasi Sms Spam Menggunakan Support Vector Machine," *Jurnal Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 275–280, Sep. 2019, doi: 10.33480/pilar.v15i2.693.
[11] D. Kawade and K. Oza, "Content-Based SMS Spam Filtering Using Machine Learning Technique," *International Journal of Computer Engineering and Applications*, vol. 13, no. 4, 2018.
[12] M. Bassiouni, M. Ali, and E. A. El-Dahshan, "Ham and Spam E-Mails Classification Using Machine Learning Techniques," *Journal of Applied Security Research*, vol. 13, no. 3, pp. 315–331, 2018, doi: 10.1080/19361610.2018.1463136.
[13] A. K. Jain, S. K. Yadav, and N. Choudhary, "A novel Approach to Detect Spam and Smishing SMS using Machine Learning Techniques," *International Journal of E-Services and Mobile Applications*, vol. 12, no. 1, pp. 21–38, 2020, doi: 10.4018/IJESMA.2020010102.
[14] N. K. Nagwani and A. Sharaff, "SMS Spam Filtering and Thread Identification using Bi-Level Text Classification and Clustering Techniques," *Journal of Information Science*, vol. 43, no. 1, pp. 1–13, 2017, doi: 10.1177/0165551515616310.
[15] A. Ghourabi, M. A. Mahmood, and Q. M. Alzubi, "A hybrid CNN-LSTM model for SMS spam detection in arabic and english messages," *Future Internet*, vol. 12, no. 9, pp. 1–16, 2020, doi: 10.3390/FI12090156.
[16] M. Manap, M. H. Jopri, A. R. Abdullah, R. Karim, M. R. Yusoff, and A. H. Azahar, "A verification of periodogram technique for harmonic source diagnostic analytic by using logistic regression," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 17, no. 1, pp. 497–507, 2019, doi: 10.12928/TELKOMNIKA.v17i1.10390.

Submitted: **13/04/2022**; Accepted: **17/05/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1290-1296
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4019

[17] N. Shiri Harzevili and S. H. Alizadeh, "Mixture of Latent Multinomial Naïve Bayes Classifier," *Applied Soft Computing Journal*, vol. 69, pp. 516–527, 2018, doi: 10.1016/j.asoc.2018.04.020.
[18] J. Feldman, A. Thomas-Bachli, J. Forsyth, Z. H. Patel, and K. Khan, "Development of a Global Infectious Disease Activity Database using Natural Language Processing, Machine Learning, and Human Expertise," *Journal of the American Medical Informatics Association*, vol. 26, no. 11, pp. 1355–1359, 2019, doi: 10.1093/jamia/ocz112.
[19] H. M. Safhi, B. Frikh, and B. Ouhbi, "Assessing reliability of Big Data Knowledge Discovery process," *Procedia Computer Science*, vol. 148, pp. 30–36, 2019, doi: 10.1016/j.procs.2019.01.005.
[20] X. Zheng, M. Wang, and J. Ordieres-Meré, "Comparison of Data Preprocessing Approaches for Applying Deep Learning to Human Activity Recognition in the Context of Industry 4.0," *Sensors (Switzerland)*, vol. 18, no. 7, 2018, doi: 10.3390/s18072146.
[21] S. Khomsah and Agus Sasmito Aribowo, "Model Text-Preprocessing Komentar Youtube Dalam Bahasa Indonesia," *Rekayasa Sistem dan Teknologi Informasi, RESTI*, vol. 4, no. 10, pp. 648–654, 2020.
[22] W. T. H. Putri, M. S. Prastio, R. Hendrowati, Y. Sari, and H. T. Y. Achsan, "Content-based Filtering Model for Recommendation of Indonesian Legal Article Study Case of Klinik Hukumonline," in *2019 International Workshop on Big Data and Information Security, IWBIS 2019*, 2019, pp. 9–14. doi: 10.1109/IWBIS.2019.8935726.
[23] F. Rahmi and W. Yudi, "Aplikasi SMS Spam Filtering pada Android menggunakan Naïve Bayes," Universitas Pendidikan Indonesia, 2017.
[24] S. R. Kunze and S. Auer, "Dataset retrieval," in *Proceedings - 2013 IEEE 7th International Conference on Semantic Computing, ICSC 2013*, 2013, pp. 1–8. doi: 10.1109/ICSC.2013.12.
[25] S. Vijayarani and J. Rajaraman, "Text Mining: open Source Tokenization Tools – An Analysis," *Advanced Computational Intelligence: An International Journal (ACII)*, vol. 3, no. 1, pp. 37–47, 2016, doi: 10.5121/acii.2016.3104.
[26] C. C. Aggarwal, *Machine Learning for Text*. Yorktown Heights: Springer, 2018. doi: 10.1007/978-3-319-73531-3_10.
[27] F. Rahutomo and A. R. T. H. Ririd, "Evaluasi Daftar Stopword Bahasa Indonesia," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 6, no. 1, pp. 41–47, 2019, doi: 10.25126/jtiik.2019611226.
[28] A. F. Hidayatullah, "Pengaruh Stopword Terhadap Performa Klasifikasi Tweet Berbahasa Indonesia," *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 1, no. 1, pp. 1–4, 2016.
[29] A. B. Arifa, G. F. Fitriana, and A. R. Hasan, "Temu Kembali Informasi pada Soal Ujian dengan Rencana Pembelajaran Menggunakan Vector Space Model," *Jurnal Resti*, vol. 5, no. 1, pp. 8–12, 2021.
[30] L. A. Wirasakti, R. Permadi, A. D. Hartanto, and H. Hartatik, "Pembuatan Kata Kunci Otomatis Dalam Artikel Dengan Pemodelan Topik," *Jurnal Media Informatika Budidarma*, vol. 4, no. 1, p. 27, 2020, doi: 10.30865/mib.v4i1.1707.
[31] N. Abdulloh and A. F. Hidayatullah, "Deteksi Cyberbullying pada Cuitan Media Sosial Twitter," *Automata*, vol. Vol 1, no. 1, pp. 1–5, 2019.
[32] L. Mutawalli, M. T. A. Zaen, and W. Bagye, "Klasifikasi Teks Sosial Media Twitter Menggunakan Support Vector Machine (Studi Kasus Penusukan Wiranto)," *Jurnal Informatika dan Rekayasa Elektronik*, vol. 2, no. 2, pp. 43–51, 2019, doi: 10.36595/jire.v2i2.117.
[33] A. Santoso and G. Ariyanto, "Implementasi Deep Learning Berbasis Keras untuk Pengenalan Wajah," *Emitor*, vol. 18, no. 01, pp. 15–21, 2018, doi: 10.23917/emitor.v18i01.6235.
[34] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification," *Augmented Human Research*, vol. 5, no. 1, pp. 1–16, 2020, doi: 10.1007/s41133-020-00032-0.
[35] S. Fanissa, M. A. Fauzi, and S. Adinugroho, "Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking | Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 8, pp. 2766–2770, 2018.
[36] H. Lu, H. Xu, N. Liu, Y. Zhou, and X. Wang, "Data sanity check for deep learning systems via learnt assertions," in *ASE 2019*, 2019, pp. 1–3.
[37] E. Indrayuni, "Klasifikasi Text Mining Review Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes," *Jurnal Khatulistiwa Informatika*, vol. 7, no. 1, pp. 29–36, 2019, doi: 10.31294/jki.v7i1.1.

Submitted: **13/04/2022**; Accepted: **17/05/2022**; Published: **25/07/2022**

6% Overall Similarity

Top sources found in the following databases:

- 4% Internet database
- Crossref database
- 2% Submitted Works database
- 4% Publications database
- Crossref Posted Content database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	ejournal.nusamandiri.ac.id Internet	1%
2	Mohammad Ahsanuddin, Ali Ma'sum, Nur Anisah Ridwan. "INVESTIGA... Crossref	1%
3	beei.org Internet	<1%
4	suarabutesarko.com Internet	<1%
5	CSU, San Jose State University on 2020-05-07 Submitted works	<1%
6	University of Arizona on 2021-11-18 Submitted works	<1%
7	sersc.org Internet	<1%
8	University of Westminster on 2019-01-01 Submitted works	<1%

9	proceeding.researchsynergypress.com Internet	<1%
10	ejournal.org.cn Internet	<1%
11	Xiaoxu Liu, Haoye Lu, Amiya Nayak. "A Spam Transformer Model for S... Crossref	<1%
12	ejournals.umn.ac.id Internet	<1%

Excluded from Similarity Report

- Bibliographic material
- Cited material
- Manually excluded text blocks
- Quoted material
- Small Matches (Less then 8 words)

EXCLUDED TEXT BLOCKS

JURNAL MEDIA INFORMATIKA BUDIDARMAVolume

Sriwijaya University on 2021-10-11

Faculty of Informatics

Faisal Dharma Adhinata, Apri Junaidi. "Gender Classification on Video Using FaceNet Algorithm and Supervi...

ittelkom-pwt.ac.id

online.bpostel.com

Correspondence Author Email

ejurnal.stmik-budidarma.ac.id

Copyright © 2022, MIB, Page

ejurnal.stmik-budidarma.ac.id

Copyright © 2022, MIB, Page

ejurnal.stmik-budidarma.ac.id

M. Rifauddin and A. N. Halida, “Waspada Cybercrime dan Informasi Hoax Pada Me...

jurnal.iaii.or.id

How to Recognize and Report Spam Text Messages,” ConsumerInformation, 2020....

CSU, San Jose State University on 2020-05-07

Copyright © 2022, MIB, Page

ejurnal.stmik-budidarma.ac.id

Copyright © 2022, MIB, Page

ejurnal.stmik-budidarma.ac.id

Copyright © 2022, MIB, Page

ejurnal.stmik-budidarma.ac.id

Copyright © 2022, MIB, Page

ejurnal.stmik-budidarma.ac.id

import nltknltk.download('stopwords')from nltk.tokenize import word_tokenizefro...

University of Glasgow on 2022-03-28

Bpk/Ibu MengenaiRekening AndaTerpilih SebagaiPemenang Cek 35jtDri BNI U/Inf...

gist.github.com

wt,d = t,d

Feras A. Batarseh, Dominick Perini, Qasim Wani, Laura Freeman. "Chapter 43 Explainable Artificial Intelligen...

|

coek.info

is the total numberof words in

Indian Institute of Management on 2017-07-28

is the number of class documents

University of Computer Studies on 2022-06-28

in class

Indian Institute of Management on 2017-07-28

3.1

National College of Ireland on 2020-10-01

PAPER NAME

4019-12539-2-PB.pdf

AUTHOR

Pradana Raharja

WORD COUNT

4393 Words

CHARACTER COUNT

23558 Characters

PAGE COUNT

7 Pages

FILE SIZE

513.9KB

SUBMISSION DATE

Jul 23, 2022 9:19 PM GMT+7

REPORT DATE

Jul 23, 2022 9:20 PM GMT+7

6% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 4% Internet database
- 4% Publications database
- Crossref database
- Crossref Posted Content database
- 2% Submitted Works database

Excluded from Similarity Report

- Bibliographic material
- Quoted material
- Cited material
- Small Matches (Less then 8 words)
- Manually excluded text blocks

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1290-1296
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4019

Comparative Analysis of Multinomial Naïve Bayes and Logistic Regression Models for Prediction of SMS Spam

Pradana Ananda Raharja*, Muhammad Fajar Sidiq, Diandra Chika Fransisca

Faculty of Informatics, Informatics, Institut Teknologi Telkom Purwokerto, Banyumas, Indonesia
Email: [1,*]pradana@ittelkom-pwt.ac.id, [2]fajar@ittelkom-pwt.ac.id, [3]diandra@ittelkom-pwt.ac.id
Correspondence Author Email: pradana@ittelkom-pwt.ac.id

Abstract−This research was conducted based on a report from the United States Federal Trade Commission regarding fraud through electronic text messages via SMS that fraudsters use to manipulate potential victims. Usually, scammers spread SMS spam as an intermediary for the crime. The development of a supervised learning algorithm is applied to predict SMS spam into three categories, such as SMS spam, SMS fraud, and promotional SMS. The prediction system is dividing into several stages in the development process, including data labelling, data preprocessing, modelling, and model validation. The known accuracy based on modelling using Logistic Regression using a test size of 15% is 99%, using a test size of 20% is 99%, and using a test size of 25% is 98%. The Multinomial Naïve Bayes algorithm's accuracy with a test size of 15%, 20%, 25% is 97%. So, the SMS spam prediction approach uses the logistic regression method, which has the highest accuracy.

Keywords: Fraud; SMS Spam; Supervised Learning; Model Validation

1. INTRODUCTION

The United States Federal Trade Commission states that fraud involves sending fake text messages to trick someone into providing personal information such as passwords, account numbers, and identification numbers. Fraudsters use this information to access email or bank accounts or sell victim information to other fraudsters. Fraudsters use a variety of changing scenarios to try to get the victim's attention. Standard methods include promising gifts, gift cards, or coupons and offering low or no interest credit cards. Scammers usually send fake messages stating that they have information about the victim's account or transaction. The mode used usually says that the fraudster saw some suspicious activity on the victim's account, made a claim that there was a problem with payment information, sent fake invoices, and told the victim to contact the fraudster if the victim was going to cancel the purchase. There was even an incident where a fraudster sent a victim a fake package delivery notification[1][2]. According to the spam statistics submitted by AV-TEST, Indonesia is ranked 8th out of the world's total population in the world for global spam. The law regarding the spread of spam in Indonesia is *Undang-Undang No. 11 Tahun 2008 / Undang-Undang Informasi dan Transaksi Elektronik* (UU ITE) has not been explicitly implementing. However, sending spam can be categorized as prohibited in chapter VII article 27-34, to be precise in article 33[3][4]. Short Message Service (SMS) has developed over the decades so that it is used for business activities. SMS containing text messages is more effective than email. [5]. So that SMS is used as a tool to commit crimes and lure victims into manipulating the victim's condition [6][7].

Research conducted by Sudibyo et al. regarding the classification of spam attack attributes on email using the Decision Tree approach. Research on spam attacks with a spam dataset of 4601 records consisting of 2813 records considered spam and not spam data 278 with an initial attribute of 57 with class 1 details. One carried out three testing experiments with 30%, 50%, and 70% attribute results from unique point feature 70% better result obtained from 30% or 50% with an accuracy value of 92,469%[8]. The research conducted by Fitrian et al. aims to create an email filtering application that utilizes the naive Bayes classifier method to classify email types, including SPAM or HAM emails, and lemmatization to process words into essential words. The test results used 131 email samples, and 119 files were successfully classified correctly and while the 12 files tested got the wrong prediction value. The accuracy value obtained in this study was 90.83%[9]. Research conducted by Setiyono and Pardede investigates various data mining techniques, namely Support Vector Machine, Multinomial Naïve Bayes, and Decision Tree for automatic spam detection. Our experimental results show that the Support Vector Machine algorithm is the best of the three evaluated algorithms. Support Vector Machine reached 98.33%, while Multinomial Naïve Bayes reached 98.13% and Decision Tree with 97.10% accuracy[10]. This research was developed by evaluating the comparison of algorithms and datasets so that the aim is to compare other approaches to have a more optimum accuracy of prediction.

The development of computational methods for identifying various SMS in cyberspace requires analyse different SMS patterns[11][12]. Then make predictions against spam using processed datasets[13]. In developing a data-based SMS spam detection model, we can use techniques of machine learning. However, the prediction of SMS spam using machine learning algorithms has limitations on identifying double classification results, which means it depends on the data's characteristics[14]. Analyse several machine learning algorithms in the SMS spam detection system is to protect users from cybercrime[15]. In connection with this research, several popular machine learning classification techniques are applied, including Logistic Regression (LR) and Multinomial Naïve Bayes (MNB), to provide intelligent services in information and communication technology[16][17].

Submitted: **13/04/2022**; Accepted: **17/05/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1290-1296
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4019

The algorithm's effectiveness is tested by conducting experiments on SMS spam datasets consisting of 3 SMS categories and evaluating the algorithm's effectiveness by measuring the performance of metrics precision, recall, f1-score, and accuracy for a machine learning-based SMS spam detection model[18].

2. RESEARCH METHODOLOGY

Describe the research sequence, including research design, explain data pre-processing to process text data, make predictions using machine learning-based modelling, and model validation to determine accuracy, precision, recall, and f1-score. The explanation of the research steps is supported by references so that the explanation can be accepted scientifically. The datasets used are SMS data with various types at the data selection stage, then sorted into three data categories, including original SMS, SMS Fraud SMS, and SMS Promo. Then the pre-processing data in this study intends to process text, such as removing punctuation marks, changing to lowercase, and removing stopwords. Then the text data that has gone through the preprocessing stage is transformed into an array to be easily read by the applied algorithm. Finally, its goal is to predict text based on its category at the data mining stage. This stage aims to predict new text data not yet in the datasets. Prediction results also need to be evaluated using a confusion matrix approach to determine how accurate the method used in making predictions is. As for what needs to know that the SMS spam datasets in this study have obtained permission from previous researchers to conduct development research, using the Knowledge Discovery and Data Mining (KDD) methodology [19]. The following are the research steps carried out in extracting SMS spam text data, shown in Figure 1. The process carried out during the study consisted of the following stages.

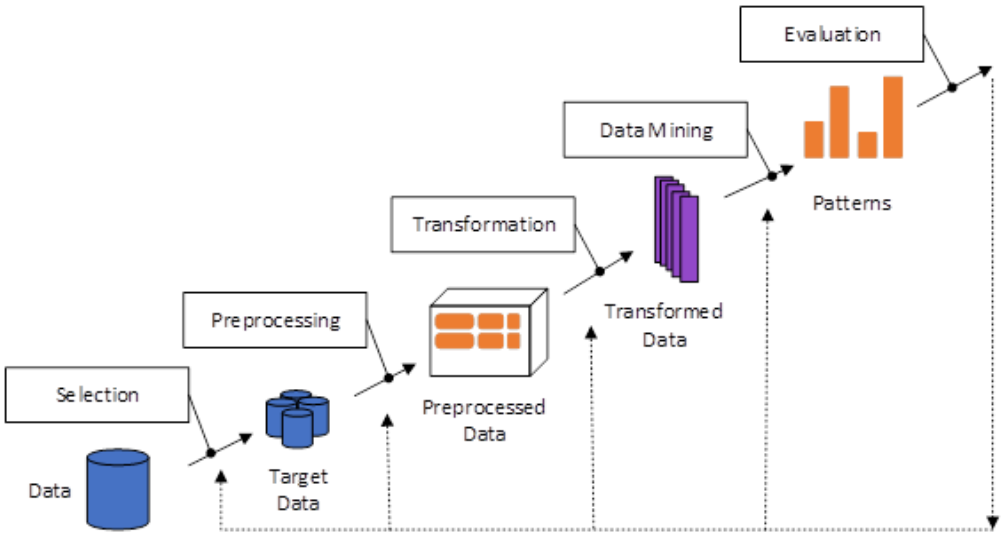


Figure 1. Data Preprocessing, Modelling, And Model Validation

2.1 Selection and Pre-processing

Selection and pre-processing are essential part of research that develops machine learning-based modelling and takes part in the analytical pipeline as our research method. The importance of applying pre-processing data in machine learning-based modelling to obtain the expected performance results[20]. The pre-processing data consisted of datasets availability, tokenization, case-folding, stop word removal, stemming, and vectorization[21][22].

a. Datasets Availability
The dataset we use in this study is SMS spam data that should make labelling by type. There are three types of SMS labels: label 0 the original SMS, label 1 is a fraud, and label 2 is SMS promotion[23]. Datasets are several datasets repositories that have information content and have relevance to research. So that data can be used to support research to be carried out[24].

b. Tokenization and Case-folding
In general, at the initial stage, the data text consists of a set of characters, and the text analysis process requires words that are available in the data set. Tokenization is simply because the text is already saved in a format that a machine can read. However, there are problems such as punctuation marks so that that punctuation marks will be removed at the tokenization stage[25]. Case-folding is briefly changing capital letters to lowercase letters to prevent ambiguity in the engine, so engine performance becomes more efficient[26].

c. Stopwords removal
One of the text processing processes in retrieving information in text or text mining or better known as stopwords removal is by deleting text from irrelevant words for indexing. There are many types of words in-

Submitted: 13/04/2022; Accepted: 17/05/2022; Published: 25/07/2022

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1290-1296
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4019

text documents, such as prepositions, conjunctions, pronouns, adjectives, Etc. Some of these words may not index the document because they are not unique or never used in the search query. Therefore, this process of filtering out words is carried out—filter by providing a stoplist list. Zipf's law is sometimes used as the basis for forming non-indexable word lists, especially in the analysis of the occurrence of words[27][28].

d. *Stemming*
The stemming process is a method for extracting a word into a root word by removing all word affixes. The prefixes include prefix, suffix, and confix[29]. The application of stemming in each language has differences depending on the morphology of each language. The result of the stemming process is stem.

2.2 Transformation

Vectorization is part of data transformation, vectorization is the last stage in pre-processing data, namely changing the form of the word represented into a number[30]. The vectorization stage uses the Term Frequency - Inverse Document Frequency (TF-IDF) method to obtain each token's weight in the vector dataset. Equation (1) is a form of the TF-IDF equation carried out on each token[31].

$$w_{t,d} = tf_{t,d} \times log \frac{N}{d_{ft}} \quad (1)$$

$tf_{t,d}$: the number of occurrences of the token *t* on the document *d.*
df_t : number of documents containing tokens *t.*
N : total documents.

2.3 Data Mining

This case study uses two-approach models as a comparison, namely LR and MNB. Modelling utilizing text classification of SMS spam is using to obtain information about fraudulent SMS messages, promo SMS messages or original SMS messages[32]. Before modelling, the datasets were testing to obtain the right level of accuracy[33]. *Logistic Regression* is a supervised learning algorithm used to classify individuals based on a logistic function. Equation (2) is an equation of LR[34].

$$ln\left(\frac{p}{1-p}\right) = B_0 + B_1X \quad (2)$$

ln : natural logarithm
B_0+B_1X : the equation known as Ordinary Least Square
P : *logistic probability*

The way MNB works is to calculate the frequency of each token appearance from the document. The document sequence of occurrences of words in the document is not to account, so the document or "*bag of word*" is processed using a multinomial distribution with equation (3)[35]. Sanity check is a testing mechanism to identify valid input data after modelling[36].

$$P(c|d) = P(c)\prod_{i=1}^{n} P(w_i|c) \quad (3)$$

$P(c|d)$: class opportunity c based on the document d, n is the total number of words in the document.

$P(c) = \frac{N_c}{N}$: opportunity class c, c is class N_c is the number of class documents c, N is the number of all documents.

$P(w_i|c) = \frac{count(w_i,c)+1}{count(c)+|V|}$: the probability of the i word in class c, $count$ (w_i, c) is the number of words to x in class c, c (c) is the total number of words in class c, $|V|$ is the number of unique words in all classes.

2.4 Evaluation

The method that is generally using calculate the accuracy in machine learning in this study is the Confusion Matrix., the Confusion Matrix loads correctly predicted classification information through the classification model. The parameters used include precision, recall, f1-score, and accuracy[37].

3. RESULT AND DISCUSSION

Based Based on the results of research conducted using methods with data pre-processing stages, modelling and model validation. The research conducted by Rami and Wibisono used SMS datasets that were label as many as 1143 messages with 569 original SMS information, 335 SMS frauds, and 239 SMS promos shown in Figure 2. The modelling applied in this study uses two supervised learning methods, namely, LR and MNB.

Submitted: **13/04/2022**; Accepted: **17/05/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1290-1296
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4019

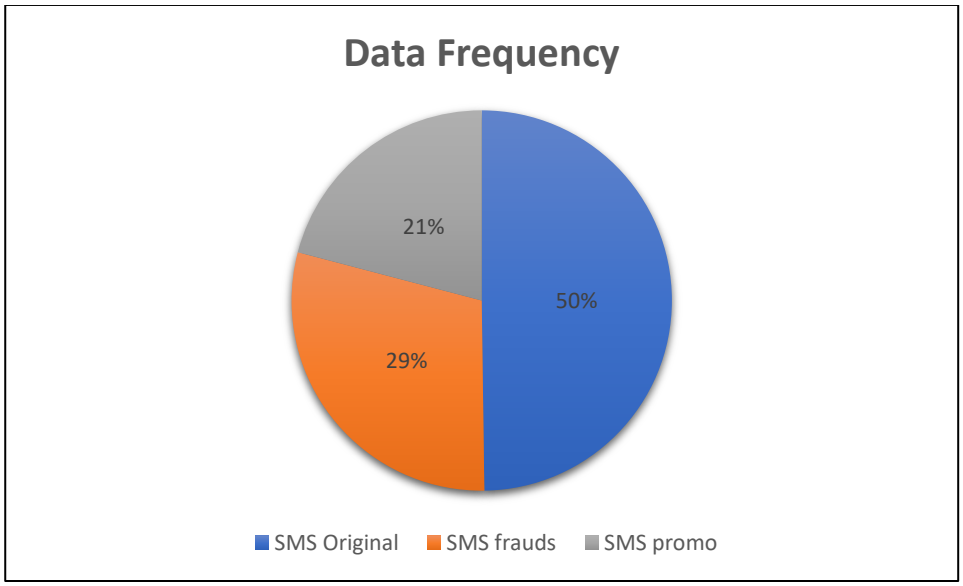


Figure 2. Datasets SMS

3.1 Selection, Pre-processing and Transformation

The data pre-processing stage consists of tokenization, case-folding, stopwords removal, stemming, and vectorization using libraries available in the Python programming language, which shown in Figure 3. Figure 4 is the output of data pre-processing which has been in the form of vectors.

```
import nltk
nltk.download('stopwords')
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from string import punctuation
sw_indo = stopwords.words("indonesian") +
list(punctuation)
```

Figure 3. Library for data pre-processing

```
array ([[0, 0, 0, ..., 0, 0, 0],
        [0, 0, 0, ..., 0, 0, 0],
        [0, 0, 0, ..., 0, 0, 0],
        ...,
        [0, 0, 0, ..., 0, 0, 0],
        [0, 0, 0, ..., 0, 0, 0],
      [0, 0, 0, ..., 0, 0, 0]])
```

Figure 4. The output of data transformation

3.2 Data mining

Prediction of modelling variation to predict three SMS text classifications using LR and MNB supported by the scikit-learn library by testing dataset sizes of 15%, 20%, and 25% of the total data and accompanied by the results of checking the accuracy of prediction algorithms, which following in Table 1.

Table 1. Results of Prediction using Logistic Regression and Multinomial Naïve Bayes

Phone Number	Sample SMS	Predictions	Method and Weight Percentage of Tested Datasets					
			Logistic Regression			Multinominal Naïve Bayes		
			15%	20%	25%	15%	20%	25%
6282299209***	Maaf Mengganggu Waktunya KAMI KOPERASI Menawarkan PNJMN-ONLINE 5jt Sampai 5OOjt Bunga 4% Pertahun Cepat & Mudah INFO WhatsApp: +6285298436***	Fraud	70,40%	56,59%	52,04%	98,78%	99,14%	99,48%
6285238123***	YTH BPK/IBU KMI MELAYANI PENGAJUAN RUPIAH CEPAT DGN PROSES CEPAT TAMPA ANGGUNAN	Fraud	99,20%	94,75%	82,12%	99,99%	99,99%	99,99%

Submitted: **13/04/2022**; Accepted: **17/05/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1290-1296
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4019

Phone Number	Sample SMS	Predictions	Method and Weight Percentage of Tested Datasets					
			Logistic Regression			Multinominal Naïve Bayes		
			15%	20%	25%	15%	20%	25%
CFC	MINIMAL 5jt-500jt INFO LENKAP HUB KMI DI WA;0823-9805-0***	Fraud	99,51%	96,22%	87,50%	100%	100%	100%
	Bpk/Ibu Mengenai Rekening Anda Terpilih Sebagai Pemenang Cek 35jt Dri BNI U/Info klik www.promobni46.tk Kode Cek 03299757							
	Hub.085288991*** DISKON 40%. 2 Ayam + 2 Chicken Strips + 2 Nasi hanya 39 RIBU NETT. Tukar SMS di CFC STASIUN PURWOKERTO hingga 14 Des. SKB. Promo *606#	Promo	96,50%	80,39%	65,30%	99,99%	99,99%	99,99%
Tokopedia	Bayar PBB gak pake antri! Cashback s.d Rp10.000 dengan kode promo: GEBYARPBB Hanya di tsel.me/pbbtokped	Promo	65,65%	51,52%	45,05%	99,99%	99,99%	99,99%
Starbucks	BELI 1 GRATIS 1. HANYA HARI INI. Semua Minuman! Tall Size! Tukarkan SMS hari ini di Starbucks terdekat (exc.Airport). S&K Berlaku. Promo*606#	Promo	99,91%	97,48%	96,79%	99,99%	99,99%	99,99%
…	…	…	…	…	…	…	…	…
085229991***	bntr lagi pulang	Original SMS	97,81%	91,01%	46,74%	99,06%	99,01%	98,82%

3.3 Evaluation

Then the accuracy performance test results by dividing the datasets sorted from lowest to highest accuracy, namely the MNB method, with datasets of 75%, 80%, and 85% having an accuracy rate of 97%. While the LR algorithm has better results, namely on datasets, 75% have an accuracy of 98%, 80% have an accuracy of 99%, and 85% have an accuracy of 99%, as shown in Figure 5.

Table 2. Evaluation of Classification Performance with Datasets Ratio

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
LR 15%	99	98	99	99
LR 20%	99	99	99	99
LR 25%	98	98	98	98
MNB 15%	97	97	97	97
MNB 20%	97	97	97	97
MNB 25%	97	97	97	97

Submitted: **13/04/2022**; Accepted: **17/05/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1290-1296
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4019

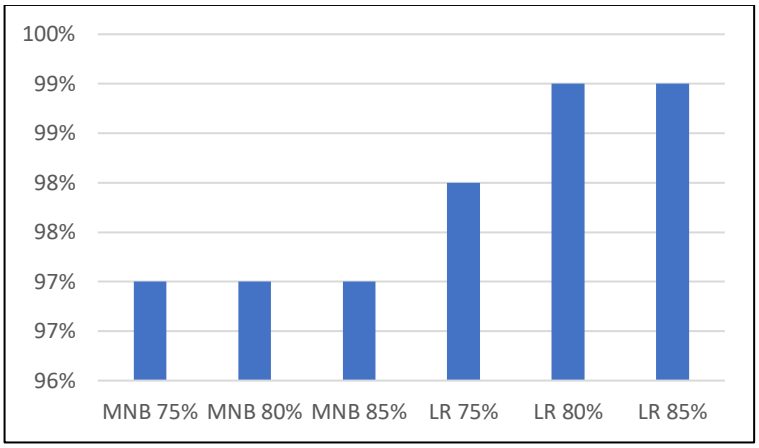


Figure 5. *Accuracy Comparison of Classification Performance*

4. CONCLUSION

Based on results of research that has been done with validation using confusion matrix, the conclusion of the LR algorithm with a test size of 15% has an accuracy of 99%, a test size of 20% has an accuracy of 99%, and a test size of 25% has an accuracy of 98%. The MNB algorithm with a test size of 15%, 20%, 25% has the same accuracy, namely 97%. With the information obtained from this study, the LR algorithm has the best accuracy in making predictions.

REFERENCES

[1] United State of America Federal Trade Commision, "How to Recognize and Report Spam Text Messages," *Consumer Information*, 2020. https://www.consumer.ftc.gov/articles/how-recognize-and-report-spam-text-messages (accessed Dec. 12, 2020).
[2] O. S. Yee, S. Sagadevan, and N. H. A. H. Malim, "Credit Card Fraud Detection Using Machine Learning As Data Mining Technique," *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 10, no. 1–4, pp. 23–27, 2018.
[3] Y. Vernanda, S. Hansun, and M. B. Kristanda, "Indonesian language email spam detection using N-gram and Naïve Bayes algorithm," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 5, pp. 2012–2019, 2020, doi: 10.11591/eei.v9i5.2444.
[4] M. Rifauddin and A. N. Halida, "Waspada Cybercrime dan Informasi Hoax Pada Media Sosial Facebook," *Khizanah al-Hikmah : Jurnal Ilmu Perpustakaan, Informasi, dan Kearsipan*, vol. 6, no. 2, pp. 98–111, 2018, doi: 10.24252/kah.v6i2a2.
[5] P. K. Roy, J. P. Singh, and S. Banerjee, "Deep learning to filter SMS Spam," *Future Generation Computer Systems*, vol. 102, pp. 524–533, 2020, doi: 10.1016/j.future.2019.09.001.
[6] I. Rahmawati, "Analisis Manajemen Resiko Ancaman Kejahatan Siber (Cyber Crime) Dalam Peningkatan Cyber Defense," *Jurnal Pertahanan & Bela Negara*, vol. 7, no. 2, pp. 51–66, 2017, doi: 10.33172/jpbh.v7i2.193.
[7] R. C. Perkins, C. J. Howell, C. E. Dodge, G. W. Burruss, and D. Maimon, "Malicious Spam Distribution: A Routine Activities Approach," *Deviant Behavior*, vol. 00, no. 00, pp. 1–17, 2020, doi: 10.1080/01639625.2020.1794269.
[8] A. Sudibyo, T. Asra, and B. Rifai, "Klasifikasi Seleksi Atribut Pada Serangan Spam Menggunakan Metode Algoritma Decision Tree," *Jurnal PILAR Nusa Mandiri*, vol. 14, no. 2, pp. 145–150, 2018, [Online]. Available: http://nusamandiri.ac.id/aji.aby@nusamandiri.ac.idhttp://bsi.ac.idhttp://nusamandiri.ac.id/
[9] H. P. Fitrian, I. Ruslianto, and R. Hidayat, "Implementasi Metode Naive Bayes Classifier Untuk Aplikasi Filtering Email Spam Dengan Lemmatization Berbasis Web," *Jurnal Coding, Sistem Komputer Untan*, vol. 06, no. 02, pp. 13–24, 2018.
[10] A. Setiyono and H. F Pardede, "Klasifikasi Sms Spam Menggunakan Support Vector Machine," *Jurnal Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 275–280, Sep. 2019, doi: 10.33480/pilar.v15i2.693.
[11] D. Kawade and K. Oza, "Content-Based SMS Spam Filtering Using Machine Learning Technique," *International Journal of Computer Engineering and Applications*, vol. 13, no. 4, 2018.
[12] M. Bassiouni, M. Ali, and E. A. El-Dahshan, "Ham and Spam E-Mails Classification Using Machine Learning Techniques," *Journal of Applied Security Research*, vol. 13, no. 3, pp. 315–331, 2018, doi: 10.1080/19361610.2018.1463136.
[13] A. K. Jain, S. K. Yadav, and N. Choudhary, "A novel Approach to Detect Spam and Smishing SMS using Machine Learning Techniques," *International Journal of E-Services and Mobile Applications*, vol. 12, no. 1, pp. 21–38, 2020, doi: 10.4018/IJESMA.2020010102.
[14] N. K. Nagwani and A. Sharaff, "SMS Spam Filtering and Thread Identification using Bi-Level Text Classification and Clustering Techniques," *Journal of Information Science*, vol. 43, no. 1, pp. 1–13, 2017, doi: 10.1177/0165551515616310.
[15] A. Ghourabi, M. A. Mahmood, and Q. M. Alzubi, "A hybrid CNN-LSTM model for SMS spam detection in arabic and english messages," *Future Internet*, vol. 12, no. 9, pp. 1–16, 2020, doi: 10.3390/FI12090156.
[16] M. Manap, M. H. Jopri, A. R. Abdullah, R. Karim, M. R. Yusoff, and A. H. Azahar, "A verification of periodogram technique for harmonic source diagnostic analytic by using logistic regression," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 17, no. 1, pp. 497–507, 2019, doi: 10.12928/TELKOMNIKA.v17i1.10390.

Submitted: **13/04/2022**; Accepted: **17/05/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1290-1296
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4019

[17] N. Shiri Harzevili and S. H. Alizadeh, "Mixture of Latent Multinomial Naïve Bayes Classifier," *Applied Soft Computing Journal*, vol. 69, pp. 516–527, 2018, doi: 10.1016/j.asoc.2018.04.020.
[18] J. Feldman, A. Thomas-Bachli, J. Forsyth, Z. H. Patel, and K. Khan, "Development of a Global Infectious Disease Activity Database using Natural Language Processing, Machine Learning, and Human Expertise," *Journal of the American Medical Informatics Association*, vol. 26, no. 11, pp. 1355–1359, 2019, doi: 10.1093/jamia/ocz112.
[19] H. M. Safhi, B. Frikh, and B. Ouhbi, "Assessing reliability of Big Data Knowledge Discovery process," *Procedia Computer Science*, vol. 148, pp. 30–36, 2019, doi: 10.1016/j.procs.2019.01.005.
[20] X. Zheng, M. Wang, and J. Ordieres-Meré, "Comparison of Data Preprocessing Approaches for Applying Deep Learning to Human Activity Recognition in the Context of Industry 4.0," *Sensors (Switzerland)*, vol. 18, no. 7, 2018, doi: 10.3390/s18072146.
[21] S. Khomsah and Agus Sasmito Aribowo, "Model Text-Preprocessing Komentar Youtube Dalam Bahasa Indonesia," *Rekayasa Sistem dan Teknologi Informasi, RESTI*, vol. 4, no. 10, pp. 648–654, 2020.
[22] W. T. H. Putri, M. S. Prastio, R. Hendrowati, Y. Sari, and H. T. Y. Achsan, "Content-based Filtering Model for Recommendation of Indonesian Legal Article Study Case of Klinik Hukumonline," in *2019 International Workshop on Big Data and Information Security, IWBIS 2019*, 2019, pp. 9–14. doi: 10.1109/IWBIS.2019.8935726.
[23] F. Rahmi and W. Yudi, "Aplikasi SMS Spam Filtering pada Android menggunakan Naïve Bayes," Universitas Pendidikan Indonesia, 2017.
[24] S. R. Kunze and S. Auer, "Dataset retrieval," in *Proceedings - 2013 IEEE 7th International Conference on Semantic Computing, ICSC 2013*, 2013, pp. 1–8. doi: 10.1109/ICSC.2013.12.
[25] S. Vijayarani and J. Rajaraman, "Text Mining: open Source Tokenization Tools – An Analysis," *Advanced Computational Intelligence: An International Journal (ACII)*, vol. 3, no. 1, pp. 37–47, 2016, doi: 10.5121/acii.2016.3104.
[26] C. C. Aggarwal, *Machine Learning for Text*. Yorktown Heights: Springer, 2018. doi: 10.1007/978-3-319-73531-3_10.
[27] F. Rahutomo and A. R. T. H. Ririd, "Evaluasi Daftar Stopword Bahasa Indonesia," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 6, no. 1, pp. 41–47, 2019, doi: 10.25126/jtiik.2019611226.
[28] A. F. Hidayatullah, "Pengaruh Stopword Terhadap Performa Klasifikasi Tweet Berbahasa Indonesia," *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 1, no. 1, pp. 1–4, 2016.
[29] A. B. Arifa, G. F. Fitriana, and A. R. Hasan, "Temu Kembali Informasi pada Soal Ujian dengan Rencana Pembelajaran Menggunakan Vector Space Model," *Jurnal Resti*, vol. 5, no. 1, pp. 8–12, 2021.
[30] L. A. Wirasakti, R. Permadi, A. D. Hartanto, and H. Hartatik, "Pembuatan Kata Kunci Otomatis Dalam Artikel Dengan Pemodelan Topik," *Jurnal Media Informatika Budidarma*, vol. 4, no. 1, p. 27, 2020, doi: 10.30865/mib.v4i1.1707.
[31] N. Abdulloh and A. F. Hidayatullah, "Deteksi Cyberbullying pada Cuitan Media Sosial Twitter," *Automata*, vol. Vol 1, no. 1, pp. 1–5, 2019.
[32] L. Mutawalli, M. T. A. Zaen, and W. Bagye, "Klasifikasi Teks Sosial Media Twitter Menggunakan Support Vector Machine (Studi Kasus Penusukan Wiranto)," *Jurnal Informatika dan Rekayasa Elektronik*, vol. 2, no. 2, pp. 43–51, 2019, doi: 10.36595/jire.v2i2.117.
[33] A. Santoso and G. Ariyanto, "Implementasi Deep Learning Berbasis Keras untuk Pengenalan Wajah," *Emitor*, vol. 18, no. 01, pp. 15–21, 2018, doi: 10.23917/emitor.v18i01.6235.
[34] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification," *Augmented Human Research*, vol. 5, no. 1, pp. 1–16, 2020, doi: 10.1007/s41133-020-00032-0.
[35] S. Fanissa, M. A. Fauzi, and S. Adinugroho, "Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking | Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 8, pp. 2766–2770, 2018.
[36] H. Lu, H. Xu, N. Liu, Y. Zhou, and X. Wang, "Data sanity check for deep learning systems via learnt assertions," in *ASE 2019*, 2019, pp. 1–3.
[37] E. Indrayuni, "Klasifikasi Text Mining Review Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes," *Jurnal Khatulistiwa Informatika*, vol. 7, no. 1, pp. 29–36, 2019, doi: 10.31294/jki.v7i1.1.

Submitted: **13/04/2022**; Accepted: **17/05/2022**; Published: **25/07/2022**

6% Overall Similarity

Top sources found in the following databases:

- 4% Internet database
- Crossref database
- 2% Submitted Works database
- 4% Publications database
- Crossref Posted Content database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	ejournal.nusamandiri.ac.id Internet	1%
2	Mohammad Ahsanuddin, Ali Ma'sum, Nur Anisah Ridwan. "INVESTIGA... Crossref	1%
3	beei.org Internet	<1%
4	suarabutesarko.com Internet	<1%
5	CSU, San Jose State University on 2020-05-07 Submitted works	<1%
6	University of Arizona on 2021-11-18 Submitted works	<1%
7	sersc.org Internet	<1%
8	University of Westminster on 2019-01-01 Submitted works	<1%

9	proceeding.researchsynergypress.com Internet	<1%
10	ejournal.org.cn Internet	<1%
11	Xiaoxu Liu, Haoye Lu, Amiya Nayak. "A Spam Transformer Model for S... Crossref	<1%
12	ejournals.umn.ac.id Internet	<1%

Excluded from Similarity Report

- Bibliographic material
- Cited material
- Manually excluded text blocks
- Quoted material
- Small Matches (Less then 8 words)

EXCLUDED TEXT BLOCKS

JURNAL MEDIA INFORMATIKA BUDIDARMAVolume

Sriwijaya University on 2021-10-11

Faculty of Informatics

Faisal Dharma Adhinata, Apri Junaidi. "Gender Classification on Video Using FaceNet Algorithm and Supervi...

ittelkom-pwt.ac.id

online.bpostel.com

Correspondence Author Email

ejurnal.stmik-budidarma.ac.id

Copyright © 2022, MIB, Page

ejurnal.stmik-budidarma.ac.id

Copyright © 2022, MIB, Page

ejurnal.stmik-budidarma.ac.id

M. Rifauddin and A. N. Halida, “Waspada Cybercrime dan Informasi Hoax Pada Me...

jurnal.iaii.or.id

How to Recognize and Report Spam Text Messages,” ConsumerInformation, 2020....

CSU, San Jose State University on 2020-05-07

Copyright © 2022, MIB, Page

ejurnal.stmik-budidarma.ac.id

Copyright © 2022, MIB, Page

ejurnal.stmik-budidarma.ac.id

Copyright © 2022, MIB, Page

ejurnal.stmik-budidarma.ac.id

Copyright © 2022, MIB, Page

ejurnal.stmik-budidarma.ac.id

import nltknltk.download('stopwords')from nltk.tokenize import word_tokenizefro...

University of Glasgow on 2022-03-28

Bpk/Ibu MengenaiRekening AndaTerpilih SebagaiPemenang Cek 35jtDri BNI U/Inf...

gist.github.com

wt,d = t,d

Feras A. Batarseh, Dominick Perini, Qasim Wani, Laura Freeman. "Chapter 43 Explainable Artificial Intelligen...

|

coek.info

is the total numberof words in

Indian Institute of Management on 2017-07-28

is the number of class documents

University of Computer Studies on 2022-06-28

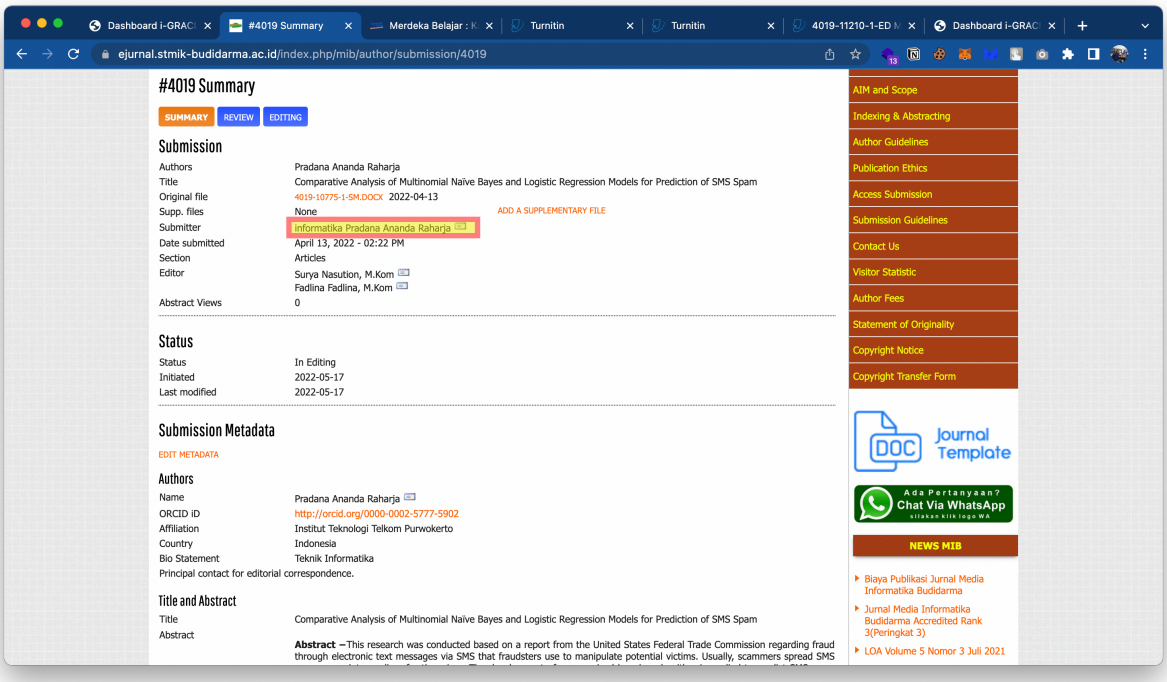
in class

Indian Institute of Management on 2017-07-28

3.1

National College of Ireland on 2020-10-01

Bukti Korespondensi



ejurnal.stmik-budidarma.ac.id/index.php/mib/author/submission/4019

#4019 Summary

SUMMARY | REVIEW | EDITING

Submission

Authors	Pradana Ananda Raharja
Title	Comparative Analysis of Multinomial Naïve Bayes and Logistic Regression Models for Prediction of SMS Spam
Original file	4019-10775-1-SM.DOCX 2022-04-13
Supp. files	None ADD A SUPPLEMENTARY FILE
Submitter	informatika Pradana Ananda Raharja
Date submitted	April 13, 2022 - 02:22 PM
Section	Articles
Editor	Surya Nasution, M.Kom Fadlina Fadlina, M.Kom
Abstract Views	0

Status

Status	In Editing
Initiated	2022-05-17
Last modified	2022-05-17

Submission Metadata

EDIT METADATA

Authors

Name	Pradana Ananda Raharja
ORCID iD	http://orcid.org/0000-0002-5777-5902
Affiliation	Institut Teknologi Telkom Purwokerto
Country	Indonesia
Bio Statement	Teknik Informatika

Principal contact for editorial correspondence.

Title and Abstract

Title	Comparative Analysis of Multinomial Naïve Bayes and Logistic Regression Models for Prediction of SMS Spam
Abstract	**Abstract** –This research was conducted based on a report from the United States Federal Trade Commission regarding fraud through electronic text messages via SMS that fraudsters use to manipulate potential victims. Usually, scammers spread SMS

AIM and Scope
Indexing & Abstracting
Author Guidelines
Publication Ethics
Access Submission
Submission Guidelines
Contact Us
Visitor Statistic
Author Fees
Statement of Originality
Copyright Notice
Copyright Transfer Form

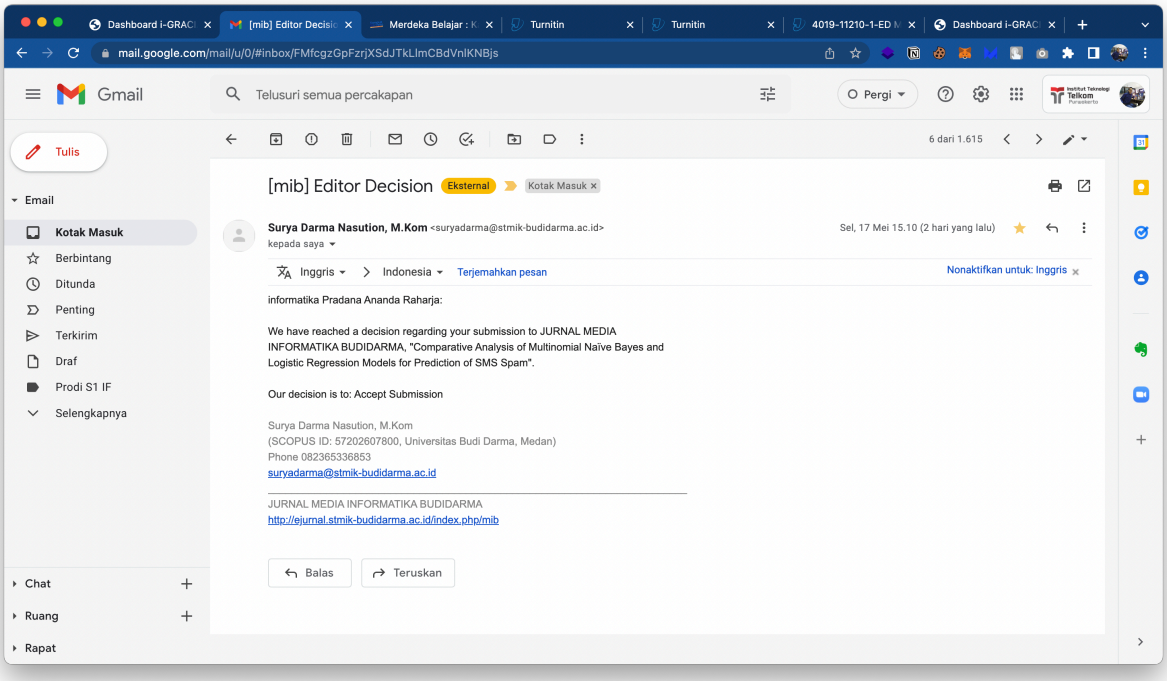
Journal Template

Ada Pertanyaan? Chat Via WhatsApp silakan klik logo WA

NEWS MIB

- Biaya Publikasi Jurnal Media Informatika Budidarma
- Jurnal Media Informatika Budidarma Accredited Rank 3(Peringkat 3)
- LOA Volume 5 Nomor 3 Juli 2021

Bukti Accepted



mail.google.com/mail/u/0/#inbox/FMfcgzGpFzrjXSdJTkLlmCBdVnlKNBjs

Gmail | Telusuri semua percakapan | Pergi

Tulis

Email
Kotak Masuk
Berbintang
Ditunda
Penting
Terkirim
Draf
Prodi S1 IF
Selengkapnya

Chat
Ruang
Rapat

6 dari 1.615

[mib] Editor Decision Eksternal Kotak Masuk ×

Surya Darma Nasution, M.Kom <suryadarma@stmik-budidarma.ac.id>
kepada saya

Sel, 17 Mei 15.10 (2 hari yang lalu)

Inggris > Indonesia Terjemahkan pesan | Nonaktifkan untuk: Inggris

informatika Pradana Ananda Raharja:

We have reached a decision regarding your submission to JURNAL MEDIA INFORMATIKA BUDIDARMA, "Comparative Analysis of Multinomial Naïve Bayes and Logistic Regression Models for Prediction of SMS Spam".

Our decision is to: Accept Submission

Surya Darma Nasution, M.Kom
(SCOPUS ID: 57202607800, Universitas Budi Darma, Medan)
Phone 082365336853
suryadarma@stmik-budidarma.ac.id

JURNAL MEDIA INFORMATIKA BUDIDARMA
http://ejurnal.stmik-budidarma.ac.id/index.php/mib

Balas | Teruskan

LoA

JURNAL MEDIA INFORMATIKA BUDIDARMA
eISSN 2548-8368 / pISSN 2614-5278
Sekretariat : UNIVERSITAS BUDI DARMA | Jl. Sisingamangaraja No. 338, Medan, Sumatera Utara
Website: https://ejurnal.stmik-budidarma.ac.id/index.php/mib
Email: mib.stmikbd@gmail.com

Medan, 10 Mei 2022

No : 538/MIB/LOA/V/2022
Lamp : -
Hal : Surat Penerimaan Naskah Publikasi Jurnal

KepadaYth,
Bapak/Ibu **Pradana Ananda Raharja**
Di Tempat

Terimakasih telah mengirimkan artikel ilmiah untuk diterbitkan pada **Jurnal Media Informatika Budidarma** (eISSN 2548-8368 / pISSN 2614-5278), dengan judul:

Comparative Analysis of Multinomial Naïve Bayes and Logistic Regression Models for Prediction of SMS Spam

Penulis: **Pradana Ananda Raharja(*), Muhammad Fajar Sidiq, Diandra Chika Fransisca**

Berdasarkan hasil review dari reviewer, artikel tersebut dinyatakan DITERIMA untuk dipublikasikan pada **Volume 6, Nomor 3, Juli 2022**.

Sebagai informasi QR-Code digunakan untuk melihat link LOA Jurnal Media Informatika Budidarma, **Volume 6, Nomor 3, Juli 2022** yang telah dikeluarkan. Mohon segera untuk mengirimkan Copyright Transfer Form ke Email Jurnal MIB.

Demikian informasi yang kami sampaikan, atas perhatiannya kami ucapkan terimakasih.

Hormat Kami,

Surya Darma Nasution, M.Kom
Ketua Editor Jurnal MIB

Tembusan:
1. Author
2. Files