SENTRI: Jurnal Riset Ilmiah

Vol.3, No.2 February 2024

ejournal.nusantaraglobal.ac.id/index.php/sentri

CLASSIFICATION OF SMS SPAM WITH N-GRAM AND PEARSON CORRELATION BASED USING MACHINE LEARNING TECHNIQUES

Nova Tri Romadloni[1], Nisa Dwi Septiyanti[2], Cucut Hariz Pratomo[3], Wakhid Kurniawan[4] , Rauhulloh Ayatulloh Khomeini Noor Bintang[5]

[1]Informatics, Faculty of Science and Technology, Universitas Muhammadiyah Karanganyar, Indonesia.
[1]Informatics, Faculty of Science and Technology, Universitas Muhammadiyah Karanganyar, Indonesia.
[1]Informatics, Faculty of Science and Technology, Universitas Muhammadiyah Karanganyar, Indonesia.
[1]Informatics, Faculty of Science and Technology, Universitas Muhammadiyah Karanganyar, Indonesia.
[1]Informatics, Faculty of Science and Technology, Universitas Muhammadiyah Karanganyar, Indonesia.
E-mail: novatrir@umuka.ac.id[1]

Article History:
Received: 20-12-2023
Revised:18-01-2024
Accepted:21-01-2024

Keywords: *Feature Selection, Machine Learning, Ngram, Pearson Correlation, SMS Classification*

Abstract*: The Short Message Service (SMS) has garnered widespread popularity due to its simplicity, reliability, and ubiquitous accessibility. This study aims to enhance the efficacy of SMS classification by refining the classification process itself. Specifically, it strives to streamline the process by diminishing feature dimensions and eliminating inconsequential attributes. The textual data undergoes preprocessing, which involves employing the N-Gram technique for feature representation, followed by meticulous feature selection utilizing Pearson Correlation. The study employs 5 of classification algorithms. Notably, the findings underscore that the optimal outcomes emerge from the fusion of the N-Gram methodology with feature selection through Pearson Correlation. Among these, the Support Vector Machine methodology stands out, exhibiting a remarkable 91.41% enhancement in accuracy without feature selection, a further improvement to 91.96% through N-Gram utilization, and a final performance of 70.80% following the inclusion of weighted correlation. However, it is imperative to acknowledge the limitations inherent in the model's generalizability, primarily stemming from the utilization of a relatively modest dataset. Despite the efficacy of Pearson correlation and N-gram-based feature selection in curbing data dimensionality and enhancing processing efficiency, certain pertinent features may have been overlooked, or the chosen attributes might not be optimally suited for specific classifications.*

INTRODUCTION

The Short Message Service (SMS) serves as a concise text messaging platform, enabling users to exchange brief messages via their mobile devices. These messages

typically span 160 characters, conveniently transmitted over cellular networks sans Internet connectivity. SMS holds a prominent role in personal interactions, connecting friends and family, while also serving as a potent tool for commercial endeavors like disseminating alerts, notifications, and marketing content to clientele. Its applicability extends across sectors, including finance, healthcare, and logistics, where it facilitates diverse functions like two-factor authentication, reminders, and seamless communication with both customers and employees[1]. Remarkably, despite the passage of time, SMS persists as a favored communication choice, valued for its simplicity, user-friendliness, and pervasive accessibility.

The intensified spotlight on SMS services has regrettably attracted malicious elements, exploiting them for their wrongful pursuits and leading to challenges for both users and service providers. SMS spam, characterized by unsolicited text messages distributed to a considerable number of mobile phone users, stands as a prevalent concern primarily harnessed for advertising or deceitful intentions [2]. Often, these messages harbor links to phishing websites or prompt recipients to divulge personal or financial information in exchange for goods or services. The perpetrators of SMS spam, frequently concealed behind anonymous or fabricated phone numbers, tend to vex and irritate the recipients [3]. In light of this, it becomes imperative for mobile users to exercise prudence when confronted with SMS messages from unfamiliar origins, abstaining from divulging any personal or financial data.

SMS or Email phishing attacks represent a crafty manifestation of social engineering, whereby perpetrators utilize text messages to manipulate individuals into revealing confidential information, encompassing passwords, bank account specifics, and personal data [4]. This art of manipulation often entails meticulously composed text messages that cunningly emulate genuine entities like banks, courier services, or social media platforms. These deceptive messages might also harbor hyperlinks redirecting unsuspecting recipients to expertly crafted fraudulent websites that exude an air of legitimacy [5]. Within these deceptive digital realms, unsuspecting victims are manipulated into divulging their login credentials or other confidential information, which then becomes accessible to the perpetrators. Furthermore, SMS phishing exhibits minimal hindrances for spammers, as they can effortlessly obtain a variety of phone numbers encompassing diverse area codes or country codes to disseminate harmful SMS messages. This ease of obtaining numbers poses a significant challenge in pinpointing and distinguishing attackers solely based on their mobile number identifiers [6]. To fortify defenses against SMS phishing attacks, it is paramount to exercise vigilance, especially in the face of unsolicited text messages, particularly those soliciting personal or sensitive information

In the domain of data mining, a multitude of methodologies have surfaced, proving efficacious in the realm of SMS spam classification. A comprehensive assessment of these methodologies who meticulously scrutinized proposed SMS spam classification techniques. In this endeavor, the author curated two distinct SMS datasets, one in Spanish encompassing 1157 legitimate messages (hams) and 199 spam messages, and another in English comprising 1119 hams and 82 spam messages. A diverse array of data mining techniques, spanning Naive Bayes, C4.5, PART, and Support Vector Machine, were harnessed for experimental exploration. The rigorous model assessment was facilitated through a robust 10-fold cross-validation methodology. The results unequivocally highlight the efficacy of employing the Naïve Bayes technique as a potent tool for proficiently classifying SMS spam. Furthermore, the study conducted by [7] aimed at implementing machine learning algorithms to discern between spam and legitimate SMS messages. A

feature set encompassing 10 distinct attributes was employed to facilitate this classification process. These attributes proved to be effective in distinguishing spam from genuine SMS messages. The utility of machine learning techniques in email spam filtering, which serves to mitigate zero-day attacks and bolster security, was established. This approach has also been extended to mobile devices to address the SMS Spam issue. However, due to the concise nature of text messages and their colloquial language, the feature set for SMS Spam differs from that used for email spam [8]. Unlike email spam, SMS messages lack graphic content and attachments, and are characterized by simplicity. The findings of this study reveal that the employed approach attained an actual positive rate of 96.5% and a remarkably low false positive rate of 1.02% through the utilization of the Random Forest classification algorithm. Extensive research has been conducted to leverage machine learning in the battle against both email and SMS spam [9], Literature review on spam content detection and classification [10]. However, these research pursuits endure, as spammers continue to adapt and refine their tactics over time. Within the domain of spam detection, numerous innovative strategies have emerged.

This refinement is anticipated to expedite classification procedures, heighten model performance, and elevate classification precision. The approach seeks to achieve these goals by enabling a more comprehensive assimilation of information from SMS texts. In pursuit of these objectives, human-categorized datasets constitute the foundation of this research. The main objective of this research is to amplify the efficacy of SMS classification by refining the very process of classification itself. Specifically, the study is geared towards optimizing the process by reducing the dimensions of features and eliminating irrelevant attributes. This enhancement is projected to streamline the classification procedure, enhance the performance of the model, and elevate the precision of classification outcomes [11]. To achieve this goal, an array of classification algorithms - including Naive Bayes, Support Vector Machines, Decision Trees, K-Nearest Neighbor, and Logistic Regression - are harnessed to effectively categorize SMS messages. The textual data undergoes meticulous preprocessing, entailing the application of the N-Gram technique for feature representation, coupled with a meticulous selection process of features guided by Pearson Correlation.

LITERATURE REVIEW

This research aims to develop an SMS classification approach using a machine learning technique that involves how we organize and apply machine algorithms to solve problems or make predictions based on data. In this approach, machine algorithms are provided with data that lacks clear labels or categories. The goal is to identify patterns or structures in the data, such as data grouping (clustering) or dimension reduction.

K-Nearest Neighbor (KNN) stands out as a widely adopted algorithm employed for classification and regression assignments within the domain of machine learning. This supervised learning approach categorizes novel data points by gauging their similarity to the training data [12]. Within KNN, the K value signifies the quantity of nearest neighbors taken into account during prediction. The algorithm computes the distance between the fresh data point and all existing training data points. Subsequently, it identifies the K closest neighbors and assigns the most prevalent class label (for classification) or computes the average value (for regression) of those neighbors to predict the label or value of the new data point [13]. A conventional logistic regression model is not applicable within a hierarchical data arrangement, where participants are grouped within clusters (cities), as this breaches the core premise of residual independence in the linear model [14]. SVM,

which stands for Support Vector Machine, is a prevalent technique in machine learning employed for classification tasks. The objective of SVM is to identify a hyperplane that optimally segregates the closest data points from each class [15]. Support Vector Machine (SVM) is a machine learning algorithm used for data classification. It analyzes and recognizes patterns in a large dataset by employing pattern recognition techniques such as statistical and mathematical methods. SVM works by inputting specific data into a system and predicting its classification. The data is then grouped into two different classes, and the new data is classified into the appropriate class [16]. Likewise with research conducted using Naive Bayes regarding sentiment classification [17].

Feature selection involves the process of cherry-picking a subset of pertinent attributes from a larger collection of attributes within a dataset. The intention is to enhance the performance of machine learning models by lessening the data's dimensionality and eliminating attributes that are insignificant or repetitive [18]. Approaches for feature selection appraise the value or pertinence of each attribute and then pick those that offer the most meaningful insights for the model. Feature selection serves as a crucial phase during the data preprocessing aspect of analyzing sentiments within beauty product reviews. It assists in identifying the most pertinent attributes (words or terms) that contribute to the categorization of sentiment. By opting for the most enlightening attributes, it's feasible to elevate the accuracy and F1-Score of the sentiment analysis model. N-gram is an n-character chunk obtained from a string. Method N-Gram is usually applied to the generation of words or characters. The use of N-Gram can also provide benefits because of the results obtained are more accurate and effective [19]. The Pearson Correlation technique is employed for determining the proximity between users by considering their ratings. The utilization of the Pearson Correlation Coefficient enhances the accuracy of computing the distances within extensive datasets [20]. Pearson Correlation method to analyze the relationship between variables obtained a large relationship simultaneously and partially between variables. Pearson Correlation is one method in statistics that can be used in analyzing the magnitude of the relationship of a variable [21].

RESEARCH METHOD

This research aims to develop an SMS classification approach using a machine learning technique that involves how we organize and apply machine algorithms to solve problems or make predictions based on data. In this approach, machine algorithms are provided with data that lacks clear labels or categories. The goal is to identify patterns or structures in the data, such as data grouping (clustering) or dimension reduction.

This research employs feature selection techniques based on Pearson Correlation and N-grams. This methodology aims to identify the most relevant features (N-grams) for classification purposes. N-grams are a natural language processing method used to understand the relationships between words in text. An N-gram refers to a sequence of N consecutive words in a text. For instance, a unigram represents a single word, a bigram consists of two consecutive words, a trigram involves three consecutive words, and so forth. This method assists in recognizing patterns and structures in text, including frequently occurring phrases or word sequences. The utilization of N-grams is beneficial for tasks such as language modeling, text classification, and other text processing tasks.

In this research, Pearson correlation analysis is employed, which is used in statistical analysis to measure the linear relationship between two variables [22]. In the context of text processing, this analysis assesses the extent to which words or features in a dataset

exhibit a linear relationship with the target variable that one seeks to predict or analyze. A high Pearson correlation (close to 1 or -1) indicates a strong relationship, while a correlation close to 0 signifies a weak or no relationship. In this case, it is crucial to ensure that these features possess a significant relationship with the appropriate category label. This methodology encompasses a series of stages, starting from data collection and extending to the evaluation of the resulting model's performance.

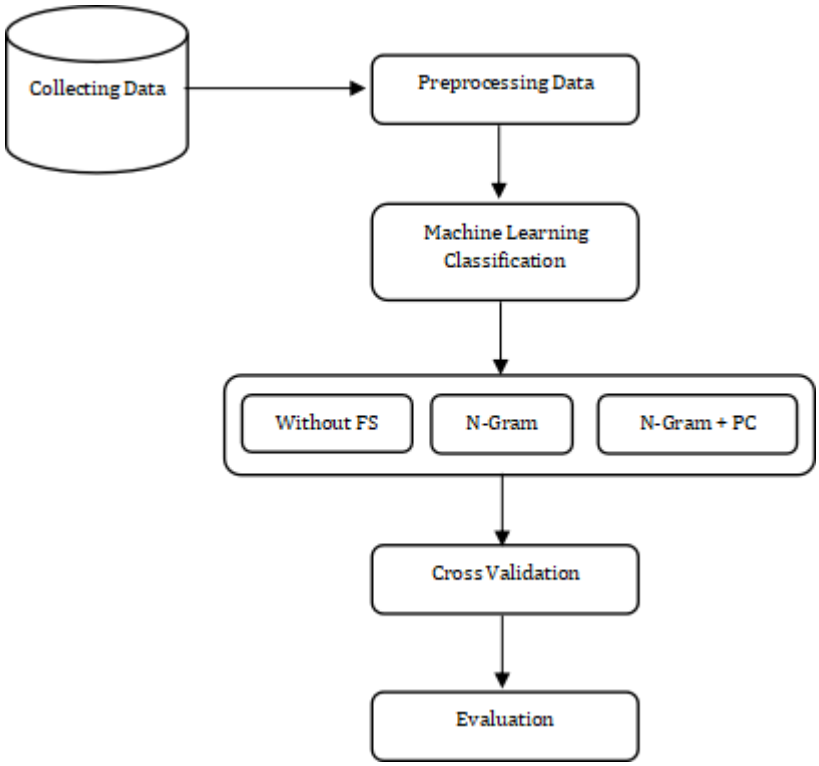

Figure 1. Research Flow

The SMS Dataset collection was gathered from relevant sources from personal individual mobile phones. Each entry in the dataset includes the text of the message and its corresponding category label, representing the context or class of the message. Subsequently, the text data undergoes a normalization process, wherein letters are converted to lowercase, punctuation and special characters are removed, and duplicate data is eliminated. Following this, the algorithmic methods are tested. Using the N-gram technique, N-gram features (such as unigrams, and bigrams) are extracted from each message in the dataset.

In this step, the construction of the classification model is carried out using the classification algorithms to be tested. Subsequently, the N-gram calculation and Pearson Correlation-based feature computation are performed on the classification model with the respective category labels. Characteristics possessing the most significant correlation values are chosen as those most inclined to exhibit a robust connection with category labels within the context of SMS classification. To evaluate the model's effectiveness, cross-validation is executed on the training data, involving the division of the dataset into numerous folds. This practice helps prevent overfitting and furnishes a more precise portrayal of the model's efficiency.

The concluding phase entails the assessment of the acquired outcomes, utilizing evaluation metrics like accuracy, precision, recall, F1-score, and AUC graphs to gauge the

model's efficacy. A comparison is drawn between the results of the model employing N-grams and Pearson Correlation-based features and the model that omits this feature selection process.

RESULT AND DISCUSSION

The research was initiated in October 2022 and data collection continued until January 2023. Data was gathered through several tens of individuals who owned mobile phones capable of receiving SMS, resulting in the acquisition of 1300 SMS messages. The collected data already had labels categorizing the SMS messages, including spam or ham. After eliminating duplicate SMS content, the dataset was reduced to 1282 messages. From this total, the messages were divided into two categories: 618 spam SMS and 664 ham SMS.

Subsequently, redundant data was removed. The following stage entailed data preprocessing, which involved identifying labels, converting numeric data to text, processing documents through tokenization, transforming letter cases, eliminating stopwords, and performing stemming using the Indonesian language dictionary. These procedures are visualized in the image provided below.

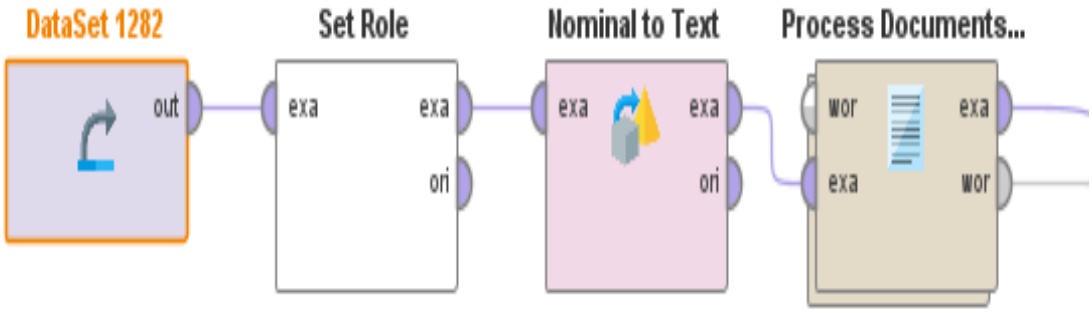

Figure 2. Data Processing

Next, The validation process is conducted using cross-validation, which employs several methods to perform tests sequentially. The first method involves utilizing a Decision Tree, followed by modifying the Naive Bayes method as the second approach, utilizing the K-Nearest Neighbor method as the third, employing Logistic Regression as the fourth, and finally, testing with the Support Vector Machines method. Each experiment or trial yields various measurements, including Accuracy, AUC (Optimistic), Precision, Recall, and F Measure. The process of testing the classification model without feature extraction yields results for each method as shown in Table 1 below.

Table. 1 Testing Without Feature Extraction

Classification Method	accuracy	AUC (optimistic)	precision	recall	F measure
Decision Tree	68.70%	0.99	62.65%	98.50%	76.56%
Naïve Bayes	87.98%	0.984	89.45%	87.20%	88.26%
K-Nearest Neighbor	90.09%	0.964	89.45%	87.20%	88.26%
Logistic Regression	90.55%	0.952	91.36%	90.51%	90.86%
Support Vector Machines	91.41%	0.956	90.93%	92.77%	91.81%

Table 1 illustrates that the Support Vector Machines classification technique yields the highest accuracy, reaching 91.41%. Testing behind are the KNN and Logistic Regression methods, displaying accuracy levels around 90%, indicating a marginal disparity. Conversely, the Decision Tree approach achieves the lowest accuracy outcome, registering at 68.70%.

After completing the testing phase without feature extraction, the next step involves testing using N-grams (bigrams). The results for each method are presented in Table 2 below.

Table. 2 Testing With N-Gram

Classification Method	accuracy	AUC (optimistic)	precision	recall	F measure
Decision Tree	68.86%	0.988	62.84%	62.84%	76.57%
Naïve Bayes	90.48%	0.991	91.66%	89.91%	90.72%
K-Nearest Neighbor	90.55%	0.967	90.07%	92.01%	90.99%
Logistic Regression	90.55%	0.967	90.07%	92.01%	90.99%
Support Vector Machines	91.96%	0.959	92.14%	92.47%	92.27%

During the examination utilizing N-Gram (bigram) analysis, the Support Vector Machines method attained the highest accuracy result at 91.96%. This result signifies an enhancement compared to the testing conducted in the absence of N-grams. However, the improvement isn't notably substantial; the difference amounts to only a few decimal points in terms of the accuracy percentage. In this assessment, a similar pattern emerges across the other four methodologies, with accuracy experiencing a modest uplift compared to the preceding testing.

After completing the series of experiments with and without N-grams, the subsequent phase entails the integration of N-grams with features derived from Pearson Correlation. In this context, the Weight By Correlation method is implemented within the dataset processing tool. The outcomes of this evaluation are presented in Table 3, displayed further below.

Table. 3 Testing With N-Gram and Features Based on Pearson Correlation

Classification Method	accuracy	AUC (optimistic)	precision	recall	F measure
Decision Tree	65.81%	0.977	93.60%	36.58%	52.33%
Naïve Bayes	69.86%	0.887	84.97%	50.88%	63.53%
K-Nearest Neighbor	69.40%	0.868	83.07%	51.49%	63.45%
Logistic Regression	74.16%	0.889	69.49%	89.45%	78.19%
Support Vector Machines	70.41%	0.873	64.67%	94.88%	76.86%

Table 3 presents the results of combining N-Gram and Pearson Correlation-based features, revealing that one of the tests conducted using the Support Vector Machines method experienced a significant decrease from 91% to 70%. This effect also influenced the performance of the other algorithm methods. Among these four methods, none of them yielded an accuracy increase after being combined with Pearson Correlation-based features. The obtained results indicate a substantial decrease, except the Decision Tree method which didn't show a significant difference in its outcomes.

Figure 3. Comparison Chart

Displayed in Figure 3 is a succinct graphical depiction that showcases the distinctions among the initial testing phase devoid of feature utilization, the trial incorporating N-Grams, and the examination involving N-Grams + PC (Pearson Correlation-based features). The outcomes portrayed in the graph elucidate that employing N-grams for testing can augment accuracy. Conversely, following the assessment involving both N-Grams and Pearson Correlation-based features, there is an observed decrease in accuracy across all classification techniques**.**

CONCLUSION

In the realm of SMS classification, the utilization of N-grams showcases an improvement in both classification effectiveness and precision. Nevertheless, upon integrating Pearson Correlation-based feature selection, there is a noticeable decline in accuracy across all methodologies, encompassing Naive Bayes, Support Vector Machines, Decision Trees, K-Nearest Neighbor, and Logistic Regression. These observations can lay the groundwork for future advancements in comprehending the significance of feature selection in textual analysis and implementing this strategy in practical scenarios spanning diverse contexts and industries.

While there are hurdles that need attention, this approach furnishes a solid foundation for the ongoing progression of natural language processing and text classification. The integration of N-grams and Pearson Correlation-based features could potentially augment the intricacy of the classification model. If the model's complexity surpasses the available data, it could result in overfitting the training data and difficulties in extrapolating to novel data instances.

REFERENCES

[1] O. Marzouk, J. Salminen, P. Zhang, and B. J. Jansen, "Which message? Which channel? Which customer? Exploring response rates in multi-channel marketing using short-form advertising," *Data Inf. Manag.*, vol. 6, no. 1, p. 100008, 2022, doi: 10.1016/j.dim.2022.100008.

[2] U. Nandagopal and S. Thirumalaivelu, "Classification of Malware with MIST and N-Gram Features Using Machine Learning," *Int. J. Intell. Eng. Syst.*, vol. 14, no. 2, pp. 323–333, 2021, doi: 10.22266/ijies2021.0430.29.

[3] C. Engineering, C. Science, and C. Science, "Mobile Sms Call Spam Filtering

Techniques," vol. 10, no. 2, pp. 112–116, 2021, doi: 10.17148/IJARCCE.2021.10217.

[4] M. Habib, H. Faris, M. A. Hassonah, J. Alqatawna, A. F. Sheta, and A. M. Al-Zoubi, "Automatic Email Spam Detection using Genetic Programming with SMOTE," *ITT 2018 - Inf. Technol. Trends Emerg. Technol. Artif. Intell.*, no. 1, pp. 185–190, 2018, doi: 10.1109/CTIT.2018.8649534.

[5] Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy," *Front. Comput. Sci.*, vol. 3, no. March, pp. 1–23, 2021, doi: 10.3389/fcomp.2021.563060.

[6] N. Choudhary and A. K. Jain, "Comparative analysis of mobile phishing detection and prevention approaches," *Smart Innov. Syst. Technol.*, vol. 83, no. Ictis 2017, pp. 349–356, 2018, doi: 10.1007/978-3-319-63673-3_43.

[7] N. Choudhary and A. K. Jain, "Towards filtering of SMS spam messages using machine learning based technique," *Commun. Comput. Inf. Sci.*, vol. 712, pp. 18–30, 2017, doi: 10.1007/978-981-10-5780-9_2.

[8] P. K. Roy, J. P. Singh, and S. Banerjee, "Deep learning to filter SMS Spam," *Futur. Gener. Comput. Syst.*, vol. 102, pp. 524–533, 2020, doi: 10.1016/j.future.2019.09.001.

[9] O. Abayomi-Alli, S. Misra, A. Abayomi-Alli, and M. Odusami, "A review of soft techniques for SMS spam classification: Methods, approaches and applications," *Eng. Appl. Artif. Intell.*, vol. 86, no. July, pp. 197–212, 2019, doi: 10.1016/j.engappai.2019.08.024.

[10] S. Kaddoura, G. Chandrasekaran, D. E. Popescu, and J. H. Duraisamy, "A systematic literature review on spam content detection and classification," *PeerJ Comput. Sci.*, vol. 8, 2022, doi: 10.7717/PEERJ-CS.830.

[11] M. A. Gohan, M. Andayani, M. Naufal, and Masliana, "Counseling on the Spread of Covid-19 Using a Participatory Action Research Approach in Responding to Hoax News on Social Media," vol. 1, pp. 66–73, 2021.

[12] R. Puspita and A. Widodo, "Perbandingan Metode KNN, Decision Tree, dan Naïve Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS," *J. Inform. Univ. Pamulang*, vol. 5, no. 4, p. 646, 2021, doi: 10.32493/informatika.v5i4.7622.

[13] Y. Deta Kirana and S. Al Faraby, "Sentiment Analysis of Beauty Product Reviews Using the K-Nearest Neighbor (KNN) and TF-IDF Methods with Chi-Square Feature Selection," *Open Access J Data Sci Appl*, vol. 4, no. 1, pp. 31–042, 2021, doi: 10.34818/JDSA.2021.4.71.

[14] M. Hou, X. Zhou, and R. Jiang, "What Influences Family Migration Decision of China's New Generation Rural-urban Migrants? A Multilevel Logistic Regression Analysis," *J. Geogr. Res.*, vol. 5, no. 4, pp. 1–15, 2022, doi: 10.30564/jgr.v5i4.4996.

[15] N. Hafidz and D. Yanti Liliana, "Klasifikasi Sentimen pada Twitter Terhadap WHO Terkait Covid-19 Menggunakan SVM, N-Gram, PSO," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 2, pp. 213–219, 2021, doi: 10.29207/resti.v5i2.2960.

[16] A. Setiyono and H. F. Pardede, "Klasifikasi Sms Spam Menggunakan Support Vector Machine," *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 275–280, 2019, doi: 10.33480/pilar.v15i2.693.

[17] C. Villavicencio, J. J. Macrohon, X. A. Inbaraj, J. H. Jeng, and J. G. Hsieh, "Twitter sentiment analysis towards covid-19 vaccines in the Philippines using naïve bayes," *Inf.*, vol. 12, no. 5, 2021, doi: 10.3390/info12050204.

[18] S. Sheikhi, M. T. Kheirabadi, and A. Bazzazi, "An effective model for SMS spam detection using content-based features and averaged neural network," *Int. J. Eng. Trans. B Appl.*, vol. 33, no. 2, pp. 221–228, 2020, doi: 10.5829/IJE.2020.33.02B.06.
[19] N. Arifin, U. Enri, and N. Sulistiyowati, "Penerapan Algoritma Support Vector Machine (SVM) dengan TF-IDF N-Gram untuk Text Classification," *STRING (Satuan Tulisan Ris. dan Inov. Teknol.*, vol. 6, no. 2, p. 129, 2021, doi: 10.30998/string.v6i2.10133.
[20] M. M. Dewi, "Optimasi Pearson Correlation untuk Sistem Rekomendasi menggunakan Algoritma Firefly," *J. Inform.*, vol. 9, no. 1, pp. 1–5, 2022, doi: 10.31294/inf.v9i1.10209.
[21] S. S. Harahap, "Hubungan Usia, Tingkat Pendidikan, Kemampuan Bekerja, dan Masa Bekerja Terhadap Kinerja Pegawai dengan Menggunakan Metode Pearson Correlation," *J. Teknovasi*, vol. 06, no. 02, pp. 12–26, 2019.
[22] N. T. Romadloni and Hilman F Pardede, "Seleksi Fitur Berbasis Pearson Correlation Untuk Optimasi Opinion Mining Review Pelanggan," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 3, pp. 505–510, 2019, doi: 10.29207/resti.v3i3.1189.