Classification of Public Complaints Based on Text Mining Using Modified K-Nearest Neighbor, Naïve Bayes and C4.5 Algorithm

Samsul Bahri[*1], Ema Utami[2], Asro Nasiri[3]
[1,2,3]Master of Informatics Engineering, AMIKOM University Yogyakarta, Indonesia
E-mail: [*1]**sambahrie@students.amikom.ac.id**, [2]ema.u@amikom.ac.id, [3]asro@amikom.ac.id

Abstract

To improve public services, accuracy and acceleration are needed in classifying the types of complaints so that complaints can immediately get a response from the relevant regional apparatus. This public complaint data is in text form and is imbalanced in each category of regional apparatus, so we contribute to research to compare the performance of different text mining-based classification algorithms. In addition, we also tested the resampling method to overcome imbalanced data. Method This research compares several models on public complaints, which researchers have never done before. We describe several stages, data collection, preprocessing, feature extraction, and research models from relevant research. In the final stage, testing is carried out using a confusion matrix table to show accuracy, precision, recall, and f1-score. The test results show the highest value in the Naïve Bayes algorithm with the ComplementNB model without resampling data, which is 89.58% accuracy, 86.72% precision, 82.40% recall, 84.09% f1-score. However, all scores decreased when combined with SMOTE resampling of 83.66% accuracy, 67.79% precision, 80.35% recall, 71.68% f1-score. ComplementNB can be an alternative model in the classification of public complaints with imbalanced datasets.

Keywords *— Imbalanced Data, Multiclass, Resampling Method*

1. INTRODUCTION

Complaints from the public of South Tangerang City (Tangsel) are submitted to the Siaran Tangsel Application, then it will be applied to possible regional apparatus. Various complaints require accuracy and acceleration in classifying the types of complaints so that complaints can immediately get a response from the relevant regional apparatus.

The emergence of unlimited textual information, so that the public can easily submit complaints that can contain any information, causes difficulties in handling complaints [1].

Complaints on the Siaran Tangsel application are in the form of text, so that they can be resolved by text mining. Text mining is a method used for classifying text and transforming unstructured text into semi-structured text data, to find important words between the texts so that analysis can be done [2].

Previous research on data classification showed that Modified K-Nearest Neighbor (MKNN) is able to handle better accuracy than the K-Nearest Neighbor (KNN) algorithm by ignoring computerization, time efficiency, and algorithm effectiveness [3].

The Naïve Bayes algorithm can classify complaints or not on Twitter posts, and Naïve Bayes is also included in the top 10 data mining algorithms based on the International Conference on Data Mining (ICDM) paper [4]. Naïve Bayes algorithm in classifying text by providing the best accuracy value compared to KNN [5].

Algortima C4.5 can provide solutions to problems in community complaint predictions [6]. The Metropolitan Transportation Authority (MTA) customer complaint analysis in Newyork City describes how intelligent computing can be used to understand and improve public services [7].

This imbalanced public complaint data can be seen in figure 1. Imbalanced data creates problems in Machine Learning classification and predicting results becomes difficult when there is not enough data to study. Imbalanced datasets also cause Machine Learning to be confused or wrong in the classification results, minority class data are often classified as majority class, commonly called bias, to overcome this with the Synthetic Minority Over-sampling Technique (SMOTE) resampling method [8].

Based on previous research, we contribute to the topic of public complaints research by suggesting and comparing models that can be used for the automatic classification of complaints on imbalanced data, and see if there is an effect from applying the resampling method to the model we use.

This paper is structured as follows: the first part discusses the background of the problem regarding data and literature review on public complaints. After that, we discussed the research method to be carried out. Finally, we discuss the results and conclusions of the study.

2. RESEARCH METHOD

Our research approach is to compare several models on the topic of public complaints, which researchers have never done before. We describe several stages, data collection, preprocessing, feature extraction, and research models taken from several studies related to text mining-based public complaints, for more details, see figure 1 below.

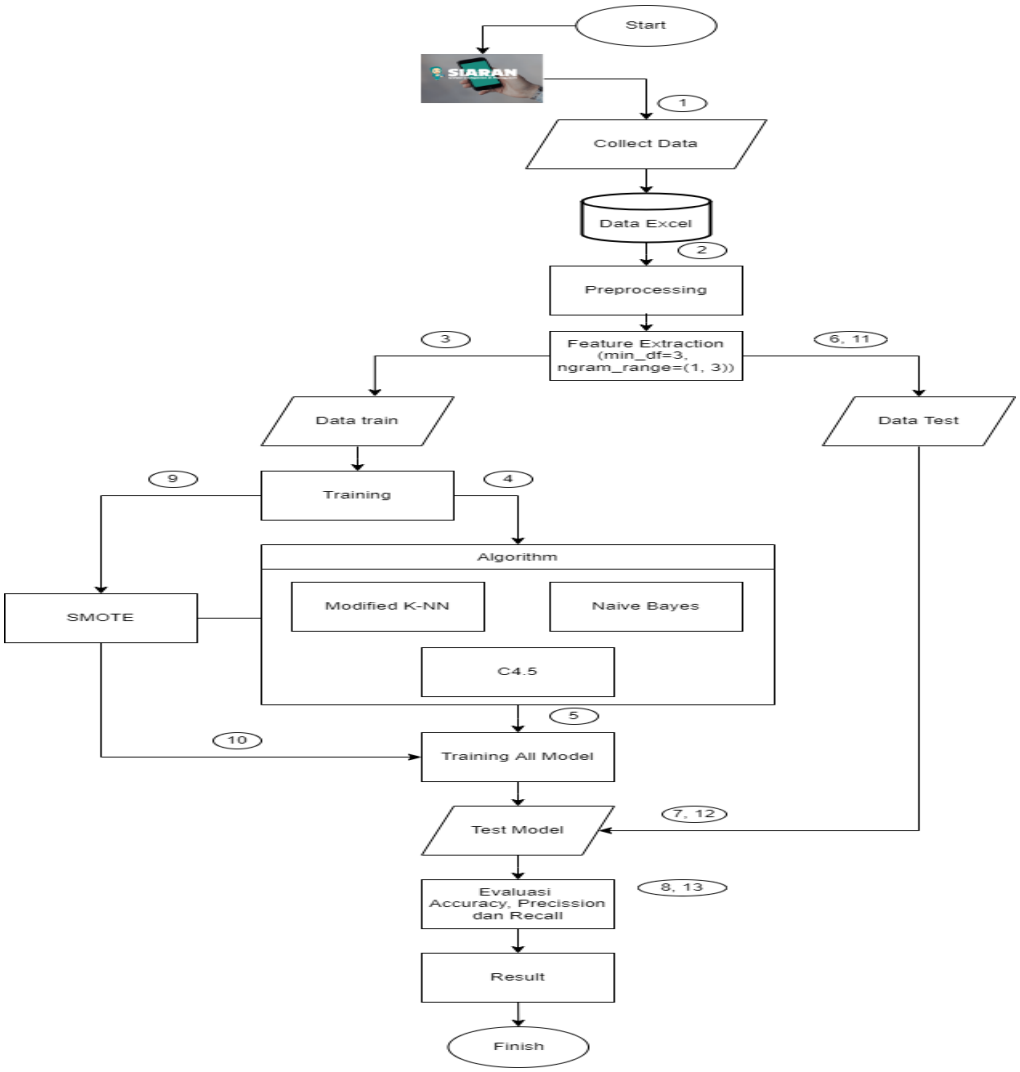

Figure 1. Research Flow

2.1. Collect Data

Public complaint data was taken from the Siaran Tangsel application from March 2017 to September 2021 and has been labeled with a total of 1181 with 11 categories of complaints based on regional apparatus as shown in Figure 2.

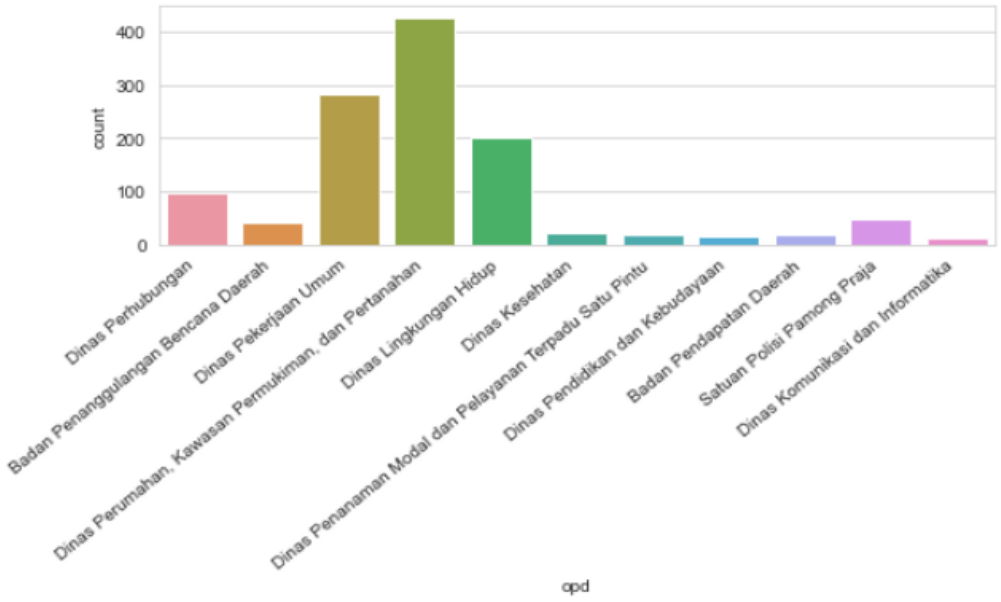

Figure 2. Distribution of Complaints to Each Regional Apparatus

2.2. Preprocessing

This stage is the initial stage in text mining-based classification. The stages in preprocessing consist of Case Folding, namely all letters in text data are converted to lowercase or lowercase, as well as deleting characters such as emoticons, numerics, urls that have no meaning, Tokenization of separating words in a sentence, Filtering these stages existing tokens in the stopword database in the literary library it will be deleted, and tokens that don't exist yet will be processed. Before stemming is done, the researcher also normalizes words that can be used to homogenize words that have the same meaning but are written differently (abbreviations, slang words, typo words/ typing error), Stemming removes the prefix and suffix for each word, and the last step is Join Text, which is cleaning the results of the stemming process.

Table 1. Result Preprocessin

Stages	Result
Raw Data	Mohon diperhatikan hampir setiap turun hujan selalu banjiir dikarenakan saluran air tdk efektif, lokasi jln bhayangkara depan pusdiklat
Case folding	mohon diperhatikan hampir setiap turun hujan selalu banjiir dikarenakan saluran air tdk efektif, lokasi jln bhayangkara depan pusdiklat
Tokenization	['mohon', 'diperhatikan', 'hampir', 'setiap', 'turun', 'hujan', 'selalu', 'banjiir', 'dikarenakan', 'saluran', 'air', 'tdk', 'efektif', 'lokasi', 'jln', 'bhayangkara', 'depan', 'pusdiklat']
Filtering	['diperhatikan', 'turun', 'hujan', 'banjiir', 'saluran', 'air', 'efektif', 'lokasi', 'bhayangkara', 'pusdiklat']
Normalisasi	['diperhatikan', 'turun', 'hujan', 'banjir', 'saluran', 'air', 'efektif', 'lokasi', 'bhayangkara', 'pusdiklat']
Stemming	['perhati', 'turun', 'hujan', 'banjir', 'saluran', 'air', 'efektif', 'lokasi', 'bhayangkara', 'pusdiklat']
Join Text	perhati turun hujan banjir saluran air efektif lokasi bhayangkara pusdiklat

We also find words that are often conveyed in public complaints as shown in figure 3. The words that often appear are ‘jalan’, ‘warga’, ‘baik’ and ‘lampu’. Repairs to lights and roads dominate public complaints in South Tangerang City through the Siaran Tangsel.

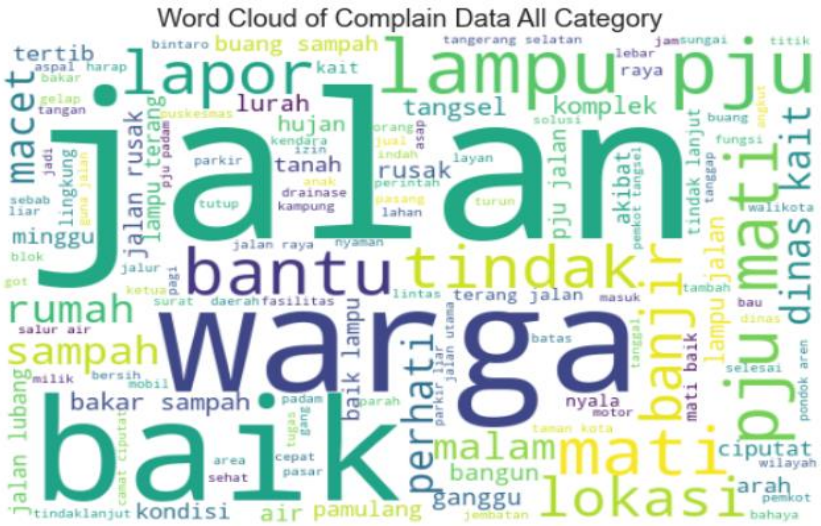

Figure 3. Words that Often Appear

2.3. Feature Extraction

Feature extraction using Term Frequency Inverse Document Frequency (TF-IDF) is a method of weighting relations between documents [9]. This method calculates the weight of term frequency (TF) which means the frequency of occurrence of a term (t) in a sentence (d) and Document Frequency (DF) which means counting the number of sentences in which a word (t) appears, the formula used :

$$TF(t,d) = \sum_{i=1}^{n} t,f \tag{1}$$

$$idf_t = log\left(\frac{N}{df_t}\right) \tag{2}$$

$$TF - IDF_{t,d} = tf_{t,d} \times idf_t \tag{3}$$

N-Gram is a sequence of N words from the sequence of extracted and selected words [10], where n is represented starting from 1. when n is 1, it is called a unigram; when n is 2, it is called a bigram; and when n is 3, it is called a trigram. In this study we used n is 3, which is called a trigram. The trigram word series will later be used as terms in the TF-IDF.

2.4. Research Model

This research can be said to be a multiclass text classification with imbalanced class data, which will affect the performance of the algorithm, for that it is important to train and test several algorithms to get the highest performance on our dataset [11]. First Modified KNN [3], The MKNN algorithm is a development method from KNN, but there are several

additional processes or developments from KNN, namely calculating the distance of training data and test data, the validity of training data is used to calculate the number of points on all training data, determining voting weights using the validity of each data on training data multiplied by weight based on Euclidean distance.

The second Naive Bayes Algorithm, where we use the Multinomial Naïve Bayes Classifier [12], is designed to count the number of times a word appears in a document with a review of a term useful in determining the sentiment of the document. Bernoulli Naïve Bayes [13], to display the presence or absence of the appropriate word in the document represented by a binary feature vector. Gaussian Naïve Bayes [14], When dealing with data that is endless, such as the TF-IDF vector, the typical assumption is that the continuous values associated with each class are distributed according to a Gaussian distribution. The last is the Complement Nave Bayes [15] algorithm which is an adjustment of Multinomial Nave Bayes which uses statistics from the complement of each class to determine the size of the model.

Finally, in C4.5 [6], is the development of the decision tree algorithm, the steps in building a decision tree are selecting an attribute as the root, creating a branch for each value, dividing the cases in the branch, repeating the process for each branch until all cases in the branch have the same class. same.

We also research resampling data using SMOTE which will be combined with the algorithm we chose above. SMOTE is one of the most popular techniques used to deal with data imbalance, which helps the minority class to achieve better classifier performance [16][17].

After feature extraction, we divided the 75% train data and 25% test data, which were then modeled into two scenarios, namely the first scenario modeling with the MKNN algorithm, Naïve Bayes with ComplementNB, GaussianNB, BernoulliNB, MultinomialNB, and Decision Tree C4.5 models without resampling data and the second scenario uses data resampling with SMOTE. We measure the performance of the model with a confusion matrix.

Confusion matrix are often used in machine learning when evaluating or describing model performance in the type of supervised classification [18].

3. RESEARCH RESULTS AND DISCUSSION

We combine these two modeling scenarios with trigrams during TF-IDF feature extraction. The results of our modeling scenarios presented in the form of tables and graphs.

Based on the first scenario, As shown in Table 2 we used accuracy, precision, recall, and f1-score [19] for performance results on the classification of public complaints. By the statement [15] that the Naïve Bayes Algorithm in the ComplementNB model works better with TF-IDF feature extraction.

Table 2. Scores by Model Non SMOTE

MODEL	Accuracy	Precision	Recall	F1-Score
MKNN	76,06%	66,57%	57,40%	60,87%
ComplementNB	89,58%	86,72%	82,40%	84,09%
MultinomialNB	77,46%	39,14%	31,39%	31,30%
BernoulliNB	72,68%	34,77%	27,24%	27,59%
GaussianNB	75,77%	68,84%	52,72%	55,91%
C4.5	74,37%	48,26%	40,53%	42,97%

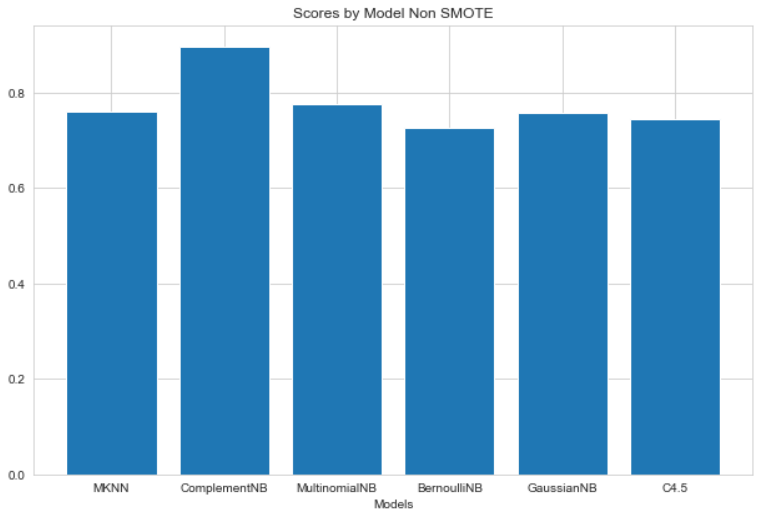

Figure 4. Score by Model Non SMOTE

Next, we oversample the data train with SMOTE to create instances in the minority class until the size of the majority class is reached [8]. Table 3 below shows the results of the model performance combined with SMOTE, it turns out that the ComplementNB model experienced a decrease in each score, but increased scores on MultinomialNB, BernoulliNB, GaussianNB, and C4.5. on Modified KNN only increases Recall and F1-Score for accuracy and precision decreases. In contrast to research [20] that the combination of SMOTE not all models of the Naïve Bayes algorithm can improve accuracy.

Table 3. Scores by Model with SMOTE

MODEL	Accuracy	Precision	Recall	F1-Score
MKNN+SMOTE	75,49%	62,70%	70,96%	64,62%
Complement+SMOTE	83,66%	67,79%	80,35%	71,68%
Multinomial+SMOTE	84,51%	76,73%	77,85%	75,42%
Bernoulli+SMOTE	78,31%	77,29%	46,11%	51,33%
Gaussian+SMOTE	76,06%	69,20%	52,82%	56,11%
C4.5+SMOTE	63,10%	60,61%	52,31%	50,99%

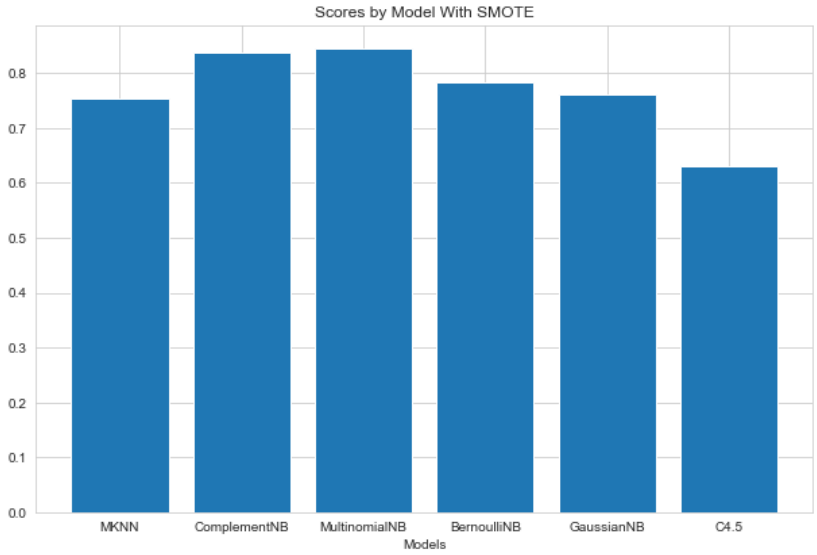

Figure 5. Scores by Model with SMOTE

4. CONCLUSION

Based on the experiments we have done, it can be said that the implementation of SMOTE resampling can improve accuracy, precision, recall, and F1-Score on the MultinomialNB, BernoulliNB, GaussianNB, and C4.5 algorithms. However, the Modified KNN algorithm only increases Recall and F1-Score, while accuracy and precision decrease. Meanwhile, ComplementNB decreased in all confusion matrix values when combined with SMOTE resampling. Although other algorithms combined with SMOTE resampling are still low below the ComplementNB algorithm without SMOTE resampling with an accuracy of 89.58%. Thus, ComplementNB can be an alternative model in the classification of public complaints with imbalanced datasets. In the future, we will try resampling methods and other classification algorithms.

5. SUGGESTED

Based on the conclusions obtained, on the topic of public complaints with data imbalanced recommended to use the ComplementNB model, or to combine it with other resampling methods such as ADASYN (Adaptive Synthetic Sampling Approach).

6. REFERENCES

[1] P. Dellia and A. Tjahyanto, "Tax Complaints Classification on Twitter Using Text Mining," *IPTEK J. Sci.*, vol. 2, no. 1, p. 11, 2017, doi: 10.12962/j23378530.v2i1.a2254.

[2] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 874, no. 1, pp. 0–6, 2020, doi: 10.1088/1757-899X/874/1/012017.

[3] Okfalisa, I. Gazalba, Mustakim, and N. G. I. Reza, "Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification," *Proc. - 2017 2nd Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2017*, vol. 2018-Janua, pp. 294–298, 2018, doi: 10.1109/ICITISEE.2017.8285514.

[4] M. Muqorobin, S. Rokhmah, I. Muslihah, and N. A. Rozaq Rais, "Classification of Community Complaints Against Public Services on Twitter," *Int. J. Comput. Inf. Syst.*, vol. 1, no. 1, pp. 7–10, 2020, doi: 10.29040/ijcis.v1i1.6.

[5] K. S. Nugroho, I. Istiadi, and F. Marisa, "Naive Bayes classifier optimization for text classification on e-government using particle swarm optimization," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 1, pp. 21–26, 2020, doi: 10.14710/jtsiskom.8.1.2020.21-26.

[6] L. N. Martin, "Comparison of C4.5 and Naïve Bayes Algorithms for Assessment of Public Complaints Services," *JITE (Journal Informatics Telecommun. Eng. Available*, vol. 3, no. 2, pp. 266–271, 2020.

[7] A. Ghazzawi and B. Alharbi, "Analysis of Customer Complaints Data using Data Mining Techniques," *Procedia Comput. Sci.*, vol. 163, pp. 62–69, 2019, doi: 10.1016/j.procs.2019.12.087.

[8] C. Padurariu and M. Elena, "Dealing with Data Imbalance in Text Classification," *Procedia Comput. Sci.*, vol. 159, pp. 736–745, 2019, doi: 10.1016/j.procs.2019.09.229.

[9] H. C. Rustamaji and O. S. Simanjuntak, "Categorical Data Classification based on Fuzzy K-Nearest Neighbor Approach," *Int. Conf. Sci. Inf. Technol.*, 2019.

[10] A. Aninditya, "Text Mining Approach Using TF-IDF and Naive Bayes for Classification of Exam Questions Based on Cognitive Level of Bloom's Taxonomy," *Proc. - 2019 IEEE Int. Conf. Internet Things Intell. Syst. IoTaIS 2019*, no. c, pp. 112–117, 2019.

[11] M. Raza, F. K. Hussain, O. K. Hussain, M. Zhao, and Z. ur Rehman, "A comparative analysis of machine learning models for quality pillar assessment of SaaS services by multi-class text classification of users' reviews," *Futur. Gener. Comput. Syst.*, vol. 101, pp. 341–371, 2019, doi: 10.1016/j.future.2019.06.022.

[12] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, "Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification," *2019 Int. Conf. Autom. Comput. Technol. Manag. ICACTM 2019*, no. April, pp. 593–596, 2019, doi: 10.1109/ICACTM.2019.8776800.

[13] G. Sanguinetti, "Text Classification using Naive Bayes = Learning only =," no. February, 2012.

[14] S. Xu, "Bayesian Naïve Bayes classifiers to text classification," *J. Inf. Sci.*, vol. 44, no. 1, pp. 48–59, 2018, doi: 10.1177/0165551516677946.

[15] T. Alsubait and D. Alfageh, "Comparison of Machine Learning Techniques for Cyberbullying Detection on YouTube Arabic Comments," *Int. J. Comput. Sci. Netw. Secur.*, vol. 21, no. 1, pp. 1–5, 2021, [Online]. Available: https://doi.org/10.22937/IJCSNS.2021.21.1.1.

[16] V. Rupapara, F. Rustam, H. F. Shahzad, A. Mehmood, I. Ashraf, and G. S. Choi, "Impact of SMOTE on Imbalanced Text Features for Toxic Comments Classification Using RVVC Model," *IEEE Access*, vol. 9, pp. 78621–78634, 2021, doi: 10.1109/ACCESS.2021.3083638.

[17] A. Mohasseb, M. Bader-El-Den, M. Cocea, and H. Liu, "Improving Imbalanced Question Classification Using Structured Smote Based Approach," *Proc. - Int. Conf. Mach. Learn. Cybern.*, vol. 2, pp. 593–597, 2018, doi: 10.1109/ICMLC.2018.8527028.

[18] O. Caelen, "A Bayesian interpretation of the confusion matrix," *Ann. Math. Artif. Intell.*, vol. 81, no. 3–4, pp. 429–450, 2017, doi: 10.1007/s10472-017-9564-8.

[19] A. I. Kadhim, "Survey on supervised machine learning techniques for automatic text classification," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 273–292, 2019, doi: 10.1007/s10462-018-09677-1.

[20] A. R. Safitri and M. A. Muslim, "Improved Accuracy of Naive Bayes Classifier for Determination of Customer Churn Uses SMOTE and Genetic Algorithms," *J. Soft Comput. Explor.*, vol. 1, no. 1, pp. 70–75, 2020.