Implementation Of Naïve Bayes Algorithm In Sentiment Analysis Of Twitter Social Media Users Regarding Their Interest To Pay The Tax

Bagas Wahyu Andrian[1], Fenina Adline Twince Tobing[2*],
Ivransa Zuhdi Pane[3], Adhi Kusnaidi[4]

[1,2,3,4] Department of Informatics, Faculty of Engineering and Informatics
Universitas Multimedia Nusantara, Tangerang, Indonesia.
[*]Corresponding Author:
Email: fenina.tobing@umn.ac.id

Abstract.

Since 2008, tax revenue has failed to reach the target set in the State Budget each year. Until 2021, tax revenue managed to reach the target that had been targeted in the 2021 state budget. In the midst of improving tax revenue, towards the end of February 2023, a case involving the son of a Directorate General of Taxes (DGT) that made the father called by the Corruption Eradication Commission (CEC) to be asked for an explanation of his assets. After the case, there were many calls in the community to stop paying taxes, which was assessed by Tauhid Ahmad as Executive Director of Indef as a form of decreased trust in tax collecting institutions. This can affect the amount of revenue from taxes because trust in the government is one of the factors that tend to affect public compliance in paying taxes. Which can affect the amount of revenue from taxes because trust in the government is one of the factors that tend to affect public compliance in paying taxes. One of the crowded calls is the pros and cons of the tax boycott movement on Twitter. With the pros and cWith the pros and cons of the movement that can affect tax revenues on Twitter social media, an assessment based on sentiment analysis is needed which is divided into positive, neutral, or negative categories. Sentiment analysis in this research is carried out using three variations of Naïve Bayes assisted by the TF-IDF word weighting model, namely Gaussian Naïve Bayes, Multinomial Naïve Bayes, and Bernoulli Naïve Bayes. Then Confussion Matrix is used to evaluate the model by obtaining the accuracy, precission, recall, and f1-score values and the use of Synthetic Minority Oversampling Technique (SMOTE) to handle unbalanced data. The results of this study on unbalanced data, the implementation of Bernoulli Naïve Bayes using the SMOTE technique on a dataset comparison of 80:20 resulted in better performance than the variations of Gaussian and Multinomial Naïve Bayes with accuracy results of 91.03%, precision, 71.11%, recall 71.43%, and f1-score of 71.18%.

Keywords: *Naïve Bayes, Sentiment Analysis, Tax Payments,and Twitter.*

I. INTRODUCTION

Indonesia is one of the countries whose most revenue comes from taxes. This can be seen from Sri Mulyani Indrawati's statement as Minister of Finance at the 2022 State Budget Realization Press Conference on January 3, 2023 which said that the 2022 State Budget was realized at IDR 2,626.4 trillion or 115.9\% of the target, where tax revenues managed to reach IDR 1,717.8 trillion or 115.6\% of the tax target and became the largest income compared to customs and excise revenues and state revenues from non-tax [1]. However, achieving the tax revenue target requires a long wait. Since 2008, tax revenue has not been able to reach the target that has been targeted in each year. Until tax revenue was able to reach the target in 2021, which was recorded at IDR 1,277.5 trillion or equivalent to 103.9\% of the tax revenue target, according to the revenue collected by the Directorate General of Taxes of the Ministry of Finance [2]. The amount of income from taxes certainly cannot be separated from the compliance of the taxpayer itself. Voluntary compliance is highly preferable as it eliminates the necessity for extensive supervision expenses while encouraging cooperative and efficient environment beneficial for both tax authorities and taxpayers. Which is the voluntary taxpayer compliance is influenced by several factors, such as trust, fairness of the tax system and the scale of power of the tax authority [3].Towards the end of February 2023, Indonesian people were shocked by the persecution case committed by the son of an official of the Directorate General of Taxes (DGT) to the son of the central board of GP Ansor. The case was first publicized on February 21, 2023 by one of the social media twitter users [4].

Along with the revelation of the persecution case, the communiity of social media users enthusiastically revealed the Tiktok account of the perpetrator who likes to show off his assets. On February 22, 2023, the identity of the perpetrator's father and his assets were revealed, which in the 2021 LHKPN reached IDR 56.1 billion. Then the father of the persecutor, who is an official of the Directorate General of

Taxes (DGT), was called by the Corruption Eradication Commission (CEC) to be asked for an explanation of these assets on March 1, 2023. The case involving this tax official led to his dismissal on March 8, 2023 and was charged with violating Law of the Republic of Indonesia Number 20 of 2001 concerning Amendments to Law Number 31 of 1999 concerning Eradication of Corruption on April 3, 2023[5].Following the case against the tax officer, a significant number of community members advocated refraining from tax payments, which was considered by Tauhid Ahmad as the Executive Director of the Institute for Developments of Economics and Finance (Indef) as a form of decreased trust in tax collection institutions [6]. One of the calls that has emerged in the community is the boycott movement to pay taxes on Twitter, because people feel that tax money is misused for the spree of unscrupulous officials or the families of tax officials themselves [7]. However, there are also many people who oppose the boycott of paying taxes because it is the same as not supporting national development in various fields. This phenomenon can threaten the tax collection institutions, considering that trust is an important factor in taxpayer compliance in paying taxes. Twitter functions as a microblog that facilitates its users to upload and share their expressions, opinions, and suggestions through 280-character messages known as "tweets", which can be factual information, opinions, expressions, sentiments, or emotions [8].

Twitter is also used by the general public to express their opinions on public topics and voice their complaints against a business or government institution. In addition, Twitter is a hub where huge amounts of data generated by users, with it being recorded that Twitter users generate 12 Gigabytes of data every day [9]. With the large amount of data on the Twitter platform, it can be sorted and used as suitable research material.Sentiment analysis is one of the research field that can be done on data spread on Twitter, which is has become an increasingly popular field of research. Sentiment analysis is one of the fields of Natural Language Processing (NLP), which is a process of automatically extracting, processing, and understanding unstructured text data to obtain information and sentiments contained in an opinion sentence that can be applied to opinions in various fields such as economics, politics, social issues, and law [10]. In conducting sentiment analysis research, there are several challenges involved such as computational cost, informal writing, language variations, and certain types of sentiment structure that occur more frequently. One particular type of sentiment structure is unstructured sentiment, which comprises informal and free-flowing writing without being restricted by any rule [11]. This is certainly related to the data found on Twitter, given that users can write and upload anything without any restrictions.Previous research involving collecting data from Twitter users conducted by Riyanto and his colleagues. In their study, they utilized a Linear Kernel Support Vector Machine as their classification model.

The results of the research indicated that the linear kernel Support Vector Machine model achieved an accuracy rate of 84.4%, precision of 86.2%, recall of 97%, and an F1-score of 88.7% [12].There is similar research regarding the application of Naive Bayes to sentiment analysis research about depression disorder involving data from Twitter. This research, which was conducted in 2023 used three variations of Naive Bayes as a classification model, namely Gaussian Naive Bayes, Multinomial Naive Bayes, and Bernoulli Naive Bayes. The research concluded that the Multinomial Naive Bayes variation yielded the highest accuracy at 90.13\%, followeed by Gaussian Naive Bayes at 88.37\%, and Bernoulli Naive Bayes in the last ranking with an accuracy score of 85.36\% [13]. In several studies, the Naive Bayes algorithm has also been observed to outperform the Support Vector Machine (SVM) algorithm. For instance, in a study conducted in 2020, the Naive Bayes algorithm demonstrated superiority over the Support Vector Machine (SVM) and K-NN algorithms, achieving an accuracy rate of 80.90\% [14].Based on the outlined problem background, a study was conducted to analyze Twitter social media users sentiments in Indonesia regarding their interest in paying taxes. This research aims to determine and find the best performance results among the three variants of the Naive Bayes method in understanding the sentiment of public opinion that is widely expressed on the Twitter platform following a case involving tax officials that has the potential to reduce taxpayers' trust in the tax collection institution. In addition, this research was conducted in the hope that it could be useful for the Indonesian government in making decisions.

II. METHODS

This research collects posts spread on the Twitter social media platform regarding Indonesian users' intentions to pay their taxes. Subsequently, the gathered data will be analyzed using three variants of Naive Bayes, and the performance of each variant will be evaluated. The various steps of this research method are elucidated in the following subsections.

A. Crawl Data

Crawl data is the first method in this research, where data will be collected from the Twitter social media platform containing predefined keywords, namely "stop bayar pajak," "berhenti bayar pajak," "stop pajak," "tetap bayar pajak," and "tetep bayar pajak" uploaded from February 21, 2023, to April 3, 2023. Data collection is automated using two scraping tools, namely Tweet-Harvest and the Snscrape library. The obtained data will be stored and merged into a Comma-Separated Values (CSV) file. The results carried out from crawling data obtained a total amount of data as much as 2617 data.

B. Labeling

After obtaining data in the data crawling step, the process continues with the labeling stage for that data. In labeling the data, there are two methods to carry out this process: manual labeling and automatic labeling. Each labeling method has its own advantages and disadvantages. Manual labeling produces more accurate data because humans can distinguish sentiments effectively. However, for a large amount of data, manual labeling can be time-consuming. On the other hand, automatic labeling has the advantage of requiring less time to label a large amount of data [15].In this study, the labeling process is conducted manually. This is because when using automatic labeling, which can only read words in English, data that is entirely in Indonesian results in many misinterpretations, leading to ambiguous labeling. Manual labeling is carried out by seeking assistance from three colleagues to understand the meaning of the tweet data and assign labels as three sentiment labels, such as positive, neutral, or negative sentiment. The final label is determined by selecting the label with the majority ratio from the three labelers, and if the label ratio is balanced, the label will be neutral. The guidelines for labelers during the data labeling process are as follows:

1) Read and understand the tweet data in the "tweet" column then label whether the tweet includes a Positive / Neutral / Negative label.
2) Positive: in the form of advice or support to keep paying taxes and disagree with the movement to boycott taxes.
3) Neutral: advertisements, news, questions, and greetings.
4) Negative: in the form of calling or agreeing to the movement that boycotts taxes.

After labeling, from 2617 data, the final labels obtained are 295 data labeled positive, 145 data labeled neutral, and 2177 data labeled neutral. This shows that the data used in this study is unbalanced where the data contains more negative sentiment. As can be seen in Fig 1.

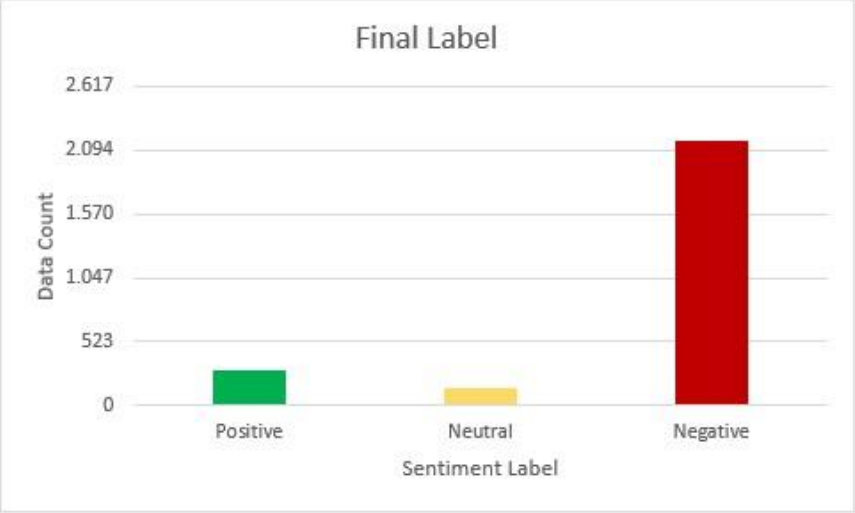

Fig 1. Final Label

C. Pre-processing

During the crawling data of tweets from the Twitter API, the obtained tweet data possesses specific features, such as emoticons, user mentions, URLs, hashtags, user mentions, and other sources of noise. This occurs due to some intrinsic characteristics of Twitter and the common usage practices of social media. Therefore, pre-processing is conducted to clean the data from noise that might potentially cover an important information [16]. The pre-processing stage is a crucial step in obtaining classification results. This is because with the use of clean and specific data, the classification results can become more accurate [17]. In brief, pre-processing is a data mining technique that involves transforming raw data into a format that is more easily understood by computers, thus addressing issues such as noise, data redundancy, and missing data [18]. The pre-processing stages used in this research are as follows:

1) Case Folding: is carried out to transform all uppercase letters in the text data into lowercase letters.
2) Cleaning: is performed to clean the data by removing unnecessary elements such as URLs, emoticons, symbols, punctuation, and usernames.
3) Tokenizing: is performed to tokenize sentences into words.
4) Normalization: is used to correct abbreviated or misspelled words into standard words.
5) Stopword: is carried out to remove words that do not carry meaningful or significant meanings, often referred to as stop words.
6) Stemming: is conducted to transform words with prefixes or suffixes into their base form

D. Data Splitting

After organizing the data during the pre-processing stage and assigning labels in the labeling stage, the process continues with the train and test data splitting phase. In this stage, the structured or organized data will be divided into training data and testing data. The data division will be performed under three comparison scenarios: 80:20, 70:30, and 60:40.

E. Feature Extraction

Feature extraction in sentiment analysis is a crucial stage. This is because feature extraction is a fundamental task in a sentiment analysis process that can directly influence the performance of sentiment classification. The purpose of this process is to extract valuable information that depicts the essential characteristics of a text [11]. In other words, feature extraction is a process of searching for and extracting features from tweets that can explain their characteristics [19].In this research, feature extraction is performed using the TF-IDF technique. The Term Frequency-Inverse Document Frequency or commonly known by the abbreviation TF-IDF is a renowned algorithm utilized to compute the weight of text and digitize the text based on the frequency and inverse document frequency of a word or phrase, known as a feature item [20]. TF-IDF is a better way to transform the textual representation of information into the Vector Space Model (VSM), where Term Frequency (TF) for a specific term (t) is computed by determining how often the term appears in a document relative to the total number of words in that document. Meanwhile, (IDF) is employed to assess the significance of the term. There have been many studies that use TF-ID as feature extraction. This is because TF-IDF successfully outperforms other methods. As a result of research conducted in 2019, TF-IDF has shown 3-4\% better results than N-gram features [21]. In addition, in a study comparing it with BM25, TF-IDF was also able to outperform BM25 technique based on the size of f1-measuer [22].

The result of the feature extraction process using TF-IDF is a vector that represents the text, and each word is assigned its respective weight [19]. The formula used to calculate the TF-IDF can be seen in the following formula [23]:

$$\mathrm{T\,F\,-\,IDF}(\mathrm{t}_k) = \mathrm{tf}_k * \log \frac{\mathrm{N}}{\mathrm{df_k}} \quad (1)$$

In formula above there are several parameters such as:

tf_k = term k frequency \newline

N = total documents \newline

$\mathrm{df_k}$ = term k document frequency

F. Apply Naïve Bayes

In this research after performing feature extraction on the three comparative scenarios, the next step is to apply the three variants of Naive Bayes for data classification. The application begins by reading one of the comparative data scenarios that has undergone feature extraction. The process then proceeds by initializing one of the Naive Bayes variants to be used. The next steps involve classifying the training data using the three Naive Bayes variants: Gaussian Naive Bayes, Multinomial Naive Bayes, and Bernoulli Naive Bayes to train these Naive Bayes variants. As a result, at the end of the process, one of the trained Naive Bayes variants is obtained and can be used to predict the test data.

- Naïve Bayes

  Naive Bayes is one of the most renowned algorithms used for sentiment analysis of text. Naive Bayes can outperform classification or machine learning algorithms such as K-Nearest Neighbor and Decision Tree. However, despite its advantages, the Naive Bayes algorithm has limitations in processing imbalanced text datasets or datasets with highly correlated features [24]. The Naive Bayes algorithm operates by forecasting future patterns derived from prior experience, which known as Bayes Theorem. The formula utilized in Bayes Theorem is as follows [25]:

$$P(C|X) = \frac{P(X|C) \times P(C)}{P(X)} \quad (2)$$

  In formula above there are several parameters such as:
  C = the target class
  X = the data
  X = the data
  P(X) = the predictor probability (prior probability)
  P(X|C) = the probability based on the conditions of the hypothesis
  P(C|X) = hypothesis probability-based on conditions (posterior probability)

- Gaussian Naïve Bayes

  Gaussian Naive Bayes is one of the variants of Naive bayes. It can be employed to calculate probabilities when the data for an attribute is continuous and consists of numeric data [25]. When dealing with continuous data, the common assumption is that continuous values correlated with each class are distributed by accessing a Gaussian distribution. In this scenario, the training data is divided based on classes, and the mean and standard deviation for each class are computed [26]. In the application of Gaussian Naive Bayes, there is a Python library that facilitates the use of this algorithm. The most well-known library for Gaussian Naive Bayes is Scikit-learn or Sklearn. The likelihood of the features is assumed to be Gaussian in this library can be observed in the formula below [27]:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \quad (3)$$

- Multinomial Naïve Bayes

  It is a fact that a term may be crucial in determining the sentiment of a document, where the multinomial model is designed to identify the term frequency, indicating how many times a term appears in a document. Additionally, the frequency of terms can assist in determining whether a term is useful for the conducted analysis or not. Therefore, Multinomial Naive Bayes becomes a suitable choice for document classification. However, this model has a drawback in which a term may sometimes appear multiple times in a document, increasing its term frequency. Simultaneously, the term may function as a stopword, lacking the potential to contribute meaning to the document while having a high term frequency. Consequently, these disruptive words need to be removed first to achieve maximum accuracy [28]. Multinomial Naive Bayes is one of the common variants of Naive Bayes extensively used in text classification, with its distribution indicated by the vector θy = (θy1 , . . . , θyn) for each class y, where n is the number of features represented in text classification, the size of the vocabulary, and θyi is the probability P(xi |y) of feature i appearing in a sample belonging to class y). The parameter θy is computed using a smoothed maximum likelihood estimation. The formula for calculating relative frequency is as follows [29]:

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_{yi} + \alpha n} \quad (4)$$

- Bernoulli Naïve Bayes

  Bernoulli Naive Bayes is also one variant of Naive Bayes, where this classification model operates efficiently with binary concepts, indicating whether an item appears or not. Bernoulli Naive Bayes takes a different approach from Multinomial Naive Bayes, as its methodology is only relevant in determining the presence of a term in the considered text [30]. The Bernoulli Naive Bayes is suitable for datasets in which features are expected to be binary, representing values as either ”True” or ”False” depending on their occurrence in a document. They can be expressed as follows [31]:

$$P(x_I|y) = P(i|y)x_i + (1 - P(i|y))(1 - x_y) \quad (5)$$

G. Synthetic Minority Oversampling Technique

In this research, we will use Synthetic Minority Oversampling or commonly abbreviated as SMOTE. This is because the data used in this study is unbalanced data. SMOTE is a statistical technique used with the aim of increasing the number of cases in the data set in a balanced way.

H. Evaluation

The last stage in this research is evaluation. Evaluation will be conducted after the three data comparison scenarios have gone through the stages of applying the three Naive Bayes variants. In this research, the evaluation is carried out using confusion matrix for the three Naive Bayes variants in all comparison scenarios using a table consisting of positive, neutral, and negative classifications. The evaluation stage using confusion matrix is carried out to obtain accuracy, precision, recall, and F1-score values to assess the performance of the three Naive Bayes variants in three different data comparison scenarios. The Confusion Matrix is composed of detailed information regarding predicted and actual values, where the evaluation of performance outcomes in a classification can be assessed using the data within this matrix. Predicted labels are represented on the X-axis, while actual labels are depicted on the Y-axis in the confusion matrix [32].

In its application to assess the performance of data classification, the confusion matrix comprises several elements as illustrated in Tabel I, where True Positive (TP) is generated when both human and method predictions are positive, and True Negative (TN) occurs when both human and method predictions are negative. Meanwhile, False Positive (FN) is used when the human prediction is positive, but the method predicts negatively, and False Positive (FP) is employed when the human prediction is negative, but the method predicts positively [33].When evaluating the accuracy of the performance results of a classification method, metrics such as accuracy, precision, and recall can be utilized as measurement values. Accuracy represents the degree of correctness between predicted values and actual values, and can be calculated using formula equation (6). Precision indicates the level of accuracy between the desired outcome by humans and the results of the system’s process, with its value determined by formula equation (7). The overall success rate of the system in the information retrieval process can be measured using recall, as outlined in formula equation (8). Precision and recall are employed to avoid measurement errors of deviation values and can be calculated using formula equation (9) [33]

Table 1. Confusion Matrix

		Real Data	
		Positive	**Negative**
Methods	**Positive**	TP	FN
	Negative	FP	TN

$$Accuracy = \frac{(T\,P + T\,N)}{(T\,P + T\,N + F\,P + F\,N)} \quad (6)$$

$$Precision = \frac{\mathrm{T\,P}}{\mathrm{T\,P\,(T\,P + F\,P)}} \quad (7)$$

$$Recall = \frac{\mathrm{T\,P}}{\mathrm{(T\,P + F\,N)}} \quad (8)$$

$$F1 - Score = \frac{(2 \times \mathrm{precision} \times \mathrm{recall})}{(\mathrm{precision} + \mathrm{recall})} \quad (9)$$

III. RESULT AND DISCUSSION

This section will explain the experimental results of the research that has been done and analyze the performance results of each naive bayes variant. Experimental results and analysis will be described based on the scenario of comparing the division of training data and test data, namely 60:40, 70:30, and 80:20.

A. 60:40 Data Comparison.

In the 60:40 dataset comparison, 2617 data that has gone through the labeling and pre-processing stages will be split into 1570 training data and 1047 test data. Where in the training data used in training the three naive bayes variants are divided into 187 with positive labels, 79 data labeled neutral, and 1304 data labeled negative. The results of the 60:40 data sharing test without the use of SMOTE can be seen in the Table II.Then the trial continued by implementing the Synthetic Minority Oversampling Technique (SMOTE) to the training data. The use of SMOTE is given a condition to increase data with positive labels to 80% of the amount of majority data or negatively labeled data where positive label data increases to 1043 data. While data with neutral labels is given a condition to increase it by 60% of the amount of negatively labeled data which increases neutral data to 782 data. Then the training data is used again to train the three variations of textitnaive bayes. The results of testing the division of 60:40 data using SMOTE can be seen in Table III.

Table 2. Comparison Results Of 60:40 Dataset Without SMOTE

Variation	Accuracy	F1-Score	Precision	Recall
GNB	35.34%	27.24%	35.64%	30.45%
MNB	83.38%	30.31%	27.79%	33.33%
BNB	84.34%	38.74%	43.29%	38.58%

Table 3. Comparison Results Of 60:40 Dataset With SMOTE

Variation	Accuracy	F1-Score	Precision	Recall
GNB	35.34%	27.15%	35.49%	30.45%
MNB	86.15%	58.88%	57.45%	61.95%
BNB	90.93%	66.97%	67.59%	67.62%

In the two tables above, you can see the performance results of each variant. The use of SMOTE can clearly improve the performance of Multinomial and Benoulli Naive Bayes significantly on accuracy, F1-Score, precision, and recall values. However, for Gaussian Naive Bayes, the best result in testing the three naive bayes variations in the 60:40 dataset comparison scenario is the bernoulli naive bayes that has used the SMOTE technique with an accuracy rate of 90.93%, f1- score of 66.97%, precision of 67.59%, and recall of 67.62%.

B. 70:30 Data Comparison

In the 70:30 dataset comparison, 2617 data points were divided into 1831 training data (70%) and 786 testing data (30%). The split training data consisted of 209 positive, 91 neutral, and 1531 negative data points, which were then used to train the three variations of Naive Bayes. The test results for the 70:30 dataset comparison trial without using the SMOTE technique can be seen in Table IV.

Table 4. Comparison Results Of 70:30 Dataset Without SMOTE

Variation	Accuracy	F1-Score	Precision	Recall
GNB	34.60%	27.19%	35.25%	30.81%
MNB	82.19%	30.07%	27.40%	33.33%
BNB	83.20%	42.14%	39.12%	39.07%

From the table above, the performance results of the three variations of Naive Bayes are relatively low. Considering that F1-Score, Precision, and Recall are also important aspects in evaluating the performance of a classification method. Therefore, the experiment was continued by using the SMOTE

technique, where positive-labeled data was increased by 80% from the majority or negative data, which increased to 1224, and neutral data was added to 918 data, equivalent to 60% of the negative data. The results of the trial on the 70:30 dataset comparison using the SMOTE technique can be seen in Table V.

Table 5. Comparison Results Of 70:30 Dataset With SMOTE

Variation	Accuracy	F1-Score	Precision	Recall
GNB	34.60%	27.19%	35.25%	30.81%
MNB	85.62%	62.17%	60.27%	65.39%
BNB	90.71%	66.41%	67.64%	67.54%

As can be compared in both Table IV and Table V, the use of the SMOTE technique improves the performance of multinomial and Bernoulli Naive Bayes. However, it cannot enhance the performance of Gaussian Naive Bayes. From the testing conducted with a 70:30 dataset split, the best results were obtained by Bernoulli Naive Bayes using the SMOTE technique with an accuracy of 90.71%

C. 80:20 Data Comparison

In the final comparison scenario, the data is split into 2093 training data and 624 test data. The training data used to train the three variations of Naive Bayes consists of 237 positively labeled data, 111 neutrally labeled, and 1745 negatively labeled data. The results of the testing in the 80:20 dataset comparison without using the SMOTE technique can be seen in Table VI.

Tbale 6. Comparison Results Of 70:30 Dataset Without SMOTE

Variation	Accuracy	F1-Score	Precision	Recall
GNB	37.02%	30.19%	38.63%	33.05%
MNB	82.44%	30.12%	27.48%	33.33%
BNB	83.59%	40.37%	42.71%	40.26%

Tbale 7. Comparison Results Of 70:30 Dataset Without SMOTE

Variation	Accuracy	F1-Score	Precision	Recall
GNB	37.02%	30.19%	38.63%	33.05%
MNB	84.92%	61.96%	59.50%	65.67%
BNB	91.03%	71.18%	71.11%	71.43%

Meanwhile, in Table VII, the results using SMOTE are presented. The use of SMOTE increases the positively labeled data to 1396. Meanwhile, for neutral data in this test, it is increased to 1047. The results are similar to the 70:30 dataset comparison experiment, where SMOTE cannot change the performance results of the Gaussian Naive Bayes variation but can improve the performance of the other two Naive Bayes variations.The optimal outcome from the 80:20 dataset split comparison testing was observed in the performance of the Bernoulli Naive Bayes employing the Synthetic Minority Over-sampling Technique (SMOTE). This is attributed to the higher values generated when compared to the results of other testing scenarios. Specifically, the utilization of Bernoulli Naive Bayes with SMOTE technique yielded an accuracy of 91.03%, an F1- score of 71.18%, precision of 71.11%, and a recall of 71.43%. Additionally, the Bernoulli Naive Bayes model with SMOTE technique in the 80:20 dataset split exhibited the highest values compared to the two other dataset split scenarios.

IV. CONCLUSION

The sentiment analysis of twitter social media sers on interest in paying taxes using naive bayes algorithm has been successfully conducted by comparing three variations of the Naive Bayes algorithm. The data utilized in this study were obtained through the crawling of Twitter data with the assistance of two scraping tools, namely tweet-harvest and snscrape. Subsequently, the data were manually labeled by three individuals, and the final labels were determined based on the majority vote. The collected data underwent preprocessing, resulting in a total of 2617 instances. The final labels were distributed as follows: 295 instances labeled as positive, 145 instances labeled as neutral, and 2177 instances labeled as negative. Based on the results of experiments conducted on imbalanced data using confusion matrix evaluation, the implementation of Bernoulli Naive Bayes with the Synthetic Minority Oversampling Technique (SMOTE) in the 80:20 dataset split demonstrated superior performance compared to the Gaussian and Multinomial Naive Bayes variations across all three dataset split scenarios. This conclusion is drawn from the achieved

performance metrics, which include an accuracy of 91.03%, an F1-score of 71.18%, precision of 71.11%, and recall of 71.43%. The spliiting of training and testing data in the three experimental scenarios indicates that higher performance is obtained with an increased amount of training data. Additionally, the application of the SMOTE technique to imbalanced data proves beneficial in enhancing the performance of Multinomial Naive Bayes and Bernoulli Naive Bayes. However, it does not contribute to the performance improvement of Gaussian Naive Bayes.

V. ACKNOWLEDGMENTS

The authors team are expresses grateful to Universitas Multimedia Nusantara for supporting this research, valueable professional support and assistance provided by Universitas Multimedia Nusantara to the author team.

REFERENCES

[1] "Minister of Finance: Exceptional state revenue performance for two consecutive years" Kemenkeu. https://www.kemenkeu.go.id/informasipublik/publikasi/berita-utama/Kinerja-Penerimaan-Negara-Luar-Biasa (accessed Oct. 30, 2023)

[2] L.J. Sembiring-Kembaren. "Finally! After 12 years of waiting, tax revenues have reached the target", CNCBIndonesia. https://www.cnbcindonesia.com/news/20220103163543-4-304218/ akhirnya-menanti-12-tahun-setoran-pajak-capai-target-juga (accessed Oct. 30, 2023)

[3] A. Djajanti. "Developing The Voluntary Taxpayer Compliance: The Scale of The Tax Authority's Power, Trust and The Fairness of The Tax System" ***Indonesian Journal of Business and Entrepreneurship (IJBE),*** Vol. 6, No. 1, pp. 86–86, Jan. 2020, DOI: http://dx.doi.org/10.17358/ IJBE.6.1.86.

[4] Y. Farouk. "Chronology of Mario Dandy, Son of Tax Official, Assaulting David Due to Romantic Issues". Suara.com. https: //www.suara.com/entertainment/2023/02/22/164009/kronologi-mariodandy-anak-pejabat-pajak-aniaya-david-gara-gara-persoalan-asmara. (accessed Oct. 30, 2023)

[5] A.Rachman. "Chronology of the RAT Case, from Wealthy Civil Servant to Incarceration by the Corruption Eradication Commission (CEC)" CNCBIndonesia. https://www.cnbcindonesia.com/news/ 20230404080107-4-427072/kronologi-kasus-rat-dari-pns-berhartajumbo-hingga-dibui-kpk (accessed Oct. 30, 2023)

[6] R.K.B. Pardede. "The Rafael Case Could Result in a Decrease in Tax Compliance" Kompas. https://www.kompas.id/baca/ekonomi/ 2023/03/02/kasus-rafael-dapat-berimbas-kepada-penurunan-kepatuhanmasyarakat-bayar-pajak?status=sukses login&%3Bstatus login= login (accessed Oct. 30, 2023)

[7] Kiwi. "Danger of the Stop Paying Taxes Hashtag, Citizens Upset about the Hedonism of Government Officials". SuaraPemred. https://www.suarapemredkalbar.com/read/ponticity/14032023/ bahaya-tagar-stop-bayar-pajak-warga-kesal-hedonisme-pejabat-negara (accessed Oct. 30, 2023).

[8] M. Ashraf et al. "Real-Time Extraction and Annotation of Social Media Contents for Predicting National Consumer Confidence Index". ***Journal of Policy Research***, Vol. 8, No. 4, pp. 292-309, Dec. 2021, DOI: https: //doi.org/10.5281/zenodo.7635142

[9] A.S. Neogi. "Sentiment analysis and classification of Indian farmers' protest using twitter data". ***International Journal of Information Management Data Insights***, Vol. 1, No. 4, pp. 100019, Nov. 2021, doi: https://doi.org/10.1016/j.jjimei.2021.100019

[10] R.Nainggolan, F.A.T. Tobing, and E.J.G.Harianja. "Analysis Sentiment in Bukalapak Comments with K-Means Clustering Method". ***IJNMT : International Journal Of New Media Technology)*** , Vol. 9, No. 2, pp. 87-92, Dec. 2022, doi: https://doi.org/10.31937/ ijnmt.v9i2.2914.

[11] M. Wankhade, A.C.S.Rao, C.Kulkarni. "A survey on sentiment analysis methods, applications, and challenges". Artifcial Intelligence Review, Vol. 55, No. 7, pp. 5731-5780, Feb. 2022, doi: https://doi.org/10.1007/ s10462-022-10144-1.

[12] Riyanto and A. Azis. "Application of the Vector Machine Support Method in Twitter Social Media Sentiment Analysis Regarding the Covid-19 Vaccine Issue in Indonesia". Journal of Applied Data Sciences, Vol. 2, No. 3, pp. 102-108, Sep. 2021, doi: https://doi.org/10.47738/ jads.v2i3.40.

[13] N.L. Lavenia and R. Permatasari. "Sentiment Analysis on Twitter Social Media Regarding Depression Disorder Using the Naive Bayes Method". ***CoreID Journal***, Vol. 1, No. 2, pp. 66-74, Jul. 2023, doi: https://doi.org/ 10.60005/coreid.v1i2.14.

[14] M. Wongkar and A. Angdresey, ”Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter,” 2019 Fourth International Conference on Informatics and Computing (ICIC), Semarang, Indonesia, 2019, pp. 1-5, doi: https://doi.org/10.1109/ICIC47613.2019.8985884.

[15] V.O. Tama, Y.Sibaroni, Adiwijaya. “Labeling Analysis in the Classification of Product Review Sentiments by using Multinomial Naive Bayes Algorithm“. ***Journal of Physics: Conference Series***, Vol. 1192, No. 1, 2019.

[16] M. Pota, M. Ventura, H. Fujita, and M. Esposito, “Multilingual evaluation of pre-processing for bert-based sentiment analysis of tweets,” Expert Systems with Applications, Vol. 181, pp. 115119, Nov. 2021.

[17] V.R. Sastri, Applications, Modern Aspects of Rare Earths and Their Complexes (Editors: V.R. Sastri, J.C. Bünzli, V. Ramachandra Rao, G.V.S. Rayudu, J.R. Perumareddi), First edition, Elsevier, 2003, pp. 893-981.

[17] M. Adnan, R. Sarno and K. R. Sungkono, ”Sentiment Analysis of Restaurant Review with Classification Approach in the Decision Tree-J48 Algorithm,” 2019 International Seminar on Application for Technology of Information and Communication (iSemantic), Semarang, Indonesia, 2019, pp. 121-126.

[18] M.R.Kurniawanda and F.A.T.Tobing.“Analysis Sentiment Cyberbullying in Instagram Comments with XGBoost Method“.***IJNMT (International Journal Of New Media Technology),***Vol.9, No.1,pp.28-34,June. 2022.

[19] R. Rahmanda and E.B. Setiawan. “Word2Vec on Sentiment Analysis with Synthetic Minority Oversampling Technique and Boosting Algorithm“. ***Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi),*** Vol. 6, No. 4, pp. 599-605, Aug. 2022, DOI: https://doi.org/10.29207/resti.v6i4.4186.

[20] T. Zhang and S.S.Ge, “An Improved TF-IDF Algorithm Based on Class Discriminative Strength for Text Categorization on Desensitized Data,“ Proceedings of the 2019 3rd international conference on innovation in artificial intelligence, Suzhou, China, 2019, pp. 39-44, doi: https: //doi.org/10.1145/3319921.3319924.

[21] R. Ahuja et al. “The Impact of Features Extraction on the Sentiment Analysis“. Procedia Computer Science, Vol. 152, pp. 341-348, 2019, DOI: https://doi.org/10.1016/j.procs.2019.05.008

[22] A.I. Kadhim, ”Term Weighting for Feature Extraction on Twitter: A Comparison Between BM25 and TF-IDF,” 2019 International Conference on Advanced Science and Engineering (ICOASE), Zakho - Duhok, Iraq, 2019, pp. 124-128, doi: https://doi.org/10.1109/ ICOASE.2019.8723825.

[23] A. Prasetyo, B. D. Septianto, G. F. Shidik and A. Z. Fanani. ”Evaluation of Feature Extraction TF-IDF in Indonesian Hoax News Classification,” 2019 International Seminar on Application for Technology of Information and Communication (iSemantic), Semarang, Indonesia, 2019, pp. 1-6.

[24] Aldinata et al. “Sentiments comparison on Twitter about LGBT“. Procedia Computer Science, Vol. 216, pp. 765-773, 2023, doi: https: //doi.org/10.1016/j.procs.2022.12.194.

[25] D. T. Barus, R. Elfarizy, F. Masri and P. H. Gunawan, ”Parallel Programming of Churn Prediction Using Gaussian Na¨ıve Bayes,” 2020 8th International Conference on Information and Communication Technology (ICoICT), Yogyakarta, Indonesia, 2020, pp. 1-4, doi: https: //doi.org/10.1109/ICoICT49345.2020.9166319.

[26] H. Kamel, D. Abdulah and J. M. Al-Tuwaijari, ”Cancer Classification Using Gaussian Naive Bayes Algorithm,” 2019 International Engineering Conference (IEC), Erbil, Iraq, 2019, pp. 165-170.

[27] V. Z. Kamila, E. Subastian and Rosmasari. ”KNN and Naive Bayes for Optional Advanced Courses Recommendation,” 2019 International Conference on Electrical, Electronics and Information Engineering (ICEEIE), Denpasar, Indonesia, 2019, pp. 306-309, doi: https://doi.org/ 10.1109/ICEEIE47180.2019.8981450

[28] G. Singh, B. Kumar, L. Gaur and A. Tyagi. ”Comparison between Multinomial and Bernoulli Na¨ıve Bayes for Text Classification,” 2019 International Conference on Automation, Computational and Technology Management (ICACTM), London, UK, 2019, pp. 593-596, doi: https://doi.org/10.1109/ICACTM.2019.8776800.

[29] M. Oljira. “Sentiment analysis of afaan oromo using machine learning approach,”. International Journal of Research Studies in Science, Engineering and Technology, vol. 7, no. 9, 2020, pp. 7–15.

[30] M. B. Ressan and R. F. Hassan, “Naive-bayes family for sentiment analysis during covid-19 pandemic and classification tweets,”. ***Indonesian Journal of Electrical Engineering and Computer Science***, vol. 28, no. 1, Okt. 2022, pp. 375.

[31] A. Kelly and M.A. Johnson, ”Investigating the Statistical Assumptions of Na¨ıve Bayes Classifiers,” 2021 55th Annual Conference on Information Sciences and Systems (CISS), Baltimore, MD, USA, 2021, pp. 1-6.

[32] P. Karthika, R. Murugeswari and R. Manoranjithem, ”Sentiment Analysis of Social Media Network Using Random Forest Algorithm,” 2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Tamilnadu, India, 2019, pp. 1-5.

[33] R. A. Laksono, K. R. Sungkono, R. Sarno and C. S. Wahyuni. ”Sentiment Analysis of Restaurant Customer Reviews on TripAdvisor using Na¨ıve Bayes,” 2019 12th International Conference on Information & Communication Technology and System (ICTS), Surabaya, Indonesia, 2019, pp. 49-54