Journal of Informatics and Data Science (J-IDS)

Vol. 03, No. 01, Juni 2024
e-ISSN: 2964-0415
https://doi.org/10.24114/j-ids.xxxxx

Sentiment Analysis of Twitter Users Regarding Taxation Topics in Indonesia Utilizing Multinomial Naive Bayes

Dewan Dinata Tarigan[1]*, Said Iskandar Al Idrus[2]
[1,2] Computer Science Department, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Medan, Indonesia

The country's income is heavily dependent on taxes, which contribute to improved public well-being. Public confidence in tax authorities plays a key role in increasing tax receipts. Therefore, it is important to measure this level of confidence. One of the methods used is sentimental analysis, which helps to understand public views on regulations, services, performance, and tax policies. One of the purposes of this study is to measure the sentiment of Twitter users towards taxation in Indonesia. Sentiment analysis involves data collection processes, initial data processing, separation of datasets, feature extraction, classification, and evaluation. The classification model used is Multinomial Naive Bayes with a comparison of 80% training data and 20% test data. The results show that 89.65% of tweets about taxation in Indonesia have negative sentiment. The model evaluation was carried out on two test scenarios, namely initial data and randomly under-sampleed data. Classification on initial data achieved accuracy of 89.97%, precision of 46.68%, and sensitivity of 33.61%. Whereas on undersampling data results, accuration reached 53.28%, accurateness of 52.66%, and sensibility of 52.52%. Analysis showed significant differences between the two scenarios in which undersammpling techniques resulted in a more balanced distribution of data. Despite this, the model still faces difficulties in classifying positive and neutral data due to the dominance of negative sentiment.

Keywords: Taxes Sentiment Analysist, Indonesian Sentimen Lexicon, Colloquial Indonesian Lexicon, TF-IDF, Multinomial Naive Bayes.

INTRODUCTION

Taxes are financial obligations that must be fulfilled by individuals or entities in accordance with the law, without direct remuneration, and are intended for the development of the country in order to improve the welfare of the people [1]. Taxes are important income for the state in improving people's welfare. Taxes are used to finance various development programs and public services, such as health, education, infrastructure, and security. In addition, taxes also serve to regulate the distribution of income and wealth in society, as well as encourage economic growth and investment[2]. Therefore, taxes have a vital role in improving public welfare, so that public compliance as taxpayers also has an important role in supporting sustainable economic growth [3].

In 2022, the Minister of Finance reported that state tax revenue increased by 34.3% on an annual basis, reaching Rp. 1,716.8 trillion [4]. This increase is inseparable from the increase in the public trust index in tax authorities in 2021. This is in line with the Slippery Slope theory which states that taxpayer compliance can be influenced by the level of public trust in the tax authority [5]. However, on February 24, 2023, a molestation case involving Mario Dandy Satriyo, the son of Directorate General of Taxes official Rafael Alun Trisambodo, raised questions about the level of public trust. Referring to the Slippery Slope theory, this event is feared to reduce the level of state revenue for taxes, because the lower the level of public trust, the lower the level of taxpayer compliance [6]. Therefore, it is important to measure this level of confidence.

The level of public trust can be measured using the sentiment analysis method [7]. In a number of studies that have been conducted, sentiment analysis tends to be applied to examine emotional responses to

[1]*Dani Dinata Tarigan.
Email addresses: dinatadewan@gmail.com (Tarigan)
DOI: 10.24114/j-ids.xxxxx

products and policies in a business context, but this is not the case in the context of government policies such as taxation [8]–[10]. Generally, the assessment of taxpayer confidence is carried out using survey instruments that allow bias to arise, because they must consider respondents' willingness and semantic differences [11], [12]. In this study, sentiment analysis is used to analyze the level of public sentiment towards taxation, so that it can contribute a more comprehensive analysis to the government, especially tax authorities. Through sentiment analysis, the government can monitor public opinion towards tax authorities, so that it can take appropriate actions to maintain public trust. Sentiment analysis in this study was conducted based on public opinion data on Twitter [13]. This is in line with the opinion of Dwi Adriansah, Country Industry Head of Twitter Indonesia, who revealed that Twitter Trends data over the past three years showed an increase in financial-related talks by 38%, including discussions about saving, investment products, crypto markets, and digital wallets.

The sentiment analysis in this study used a different approach than a number of previous studies. The main difference lies in the application of stemming, normalization, and sentiment labeling methods. Generally, the stemming, normalization, and labeling processes are done with support from TextBlob and the Natural Language Toolkit [14], [15]. However, this study applies Nazief & Andriani's algorithm for the stemming process, and refers to the Colloquial Indonesian Lexicon and Indonesian Sentiment Lexicon dictionaries for the process of normalization and labeling of sentiment. The use of nazief&andriani algorithms, as well as Colloquial Indonesian Lexicon and Indonesian Sentiment Lexicon is based on the support of features that can be used for Indonesian sentiment analysis, such as the availability of slang and polarization values of Indonesian words [16]. Unlike TextBlob and Natural Language Toolkit which are limited because they only have features and dictionaries in English. Previous research conducted by Fajar Muharram and Kana Saputra showed that the use of TextBlob and Natural Language Toolkit forced researchers to first translate words in Indonesian into English before proceeding with the sentiment analysis process, which in turn could lead to bias due to the inability to translate some slang words in Indonesian into English [17].

Sentiment analysis was conducted on 7480 tweets of Twitter users collected through scraping methods with the help of Twitter API and Snscrape. Furthermore, the TF-IDF and Naive Bayes algorithms were used to find public sentiments related to taxes. TF-IDF (Term Frequency-Inverse Document Frequency) algorithm is a statistical approach used to evaluate the significance of a word in a document. This method considers the number of words that appear in a single document and compares it to the number of words that appear in the entire document set[18]. Meanwhile, the Naive Bayes algorithm is used to classify text based on the probability of occurrence of words in a document [19]. Although there are several other algorithms that can be used in sentiment analysis, such as Random Forest, Support Vector Machine (SVM), and Neural Network [20], however, the use of TF-IDF and Naive Bayes algorithms was chosen because of its ease of implementation and effectiveness in the context of sentiment analysis. Previous research by Farah Zhafira et al. (2021) noted that the use of TF-IDF and Naive Bayes provided the highest accuracy of 97% in sentiment analysis related to the Independent Campus policy [21]. In addition, research conducted by Saragih (2021) shows that the use of Naive Bayes Classification results in an accuracy rate of 91.00% [22]. This research is limited to sentiment analysis based on the emergence and form of vocabulary, but does not delve deep into the meaning or context resulting from a series of sentences. In addition, the degree of classification accuracy is limited by the values of accuracy, precision, and sensitivity

METHODS

This research uses a quantitative approach by applying machine learning methods. This research process involves a series of stages which include the process of data collection, pre-processing of data, sharing of datasets, feature extraction, classification, and evaluation. The stages are as illustrated in Figure 1 below.

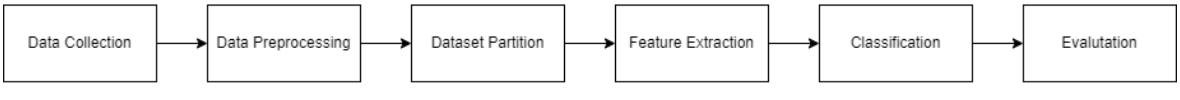

Figure 1. Research Methods

Data Collection

Data collection is executed through web scraping techniques, utilizing the Snscrape module. This scraping process is specifically conducted on the Twitter platform, targeting a set of keywords, including "Taxation," "Taxation," "Indonesian Taxation," and "Indonesian Tax." The collected tweet data spans from February 21, 2023, to March 31, 2023. As of this stage, the cumulative dataset consists of 7,480 records, which have been meticulously stored in _*.csv_ formatted documents. Examples of data samples collected can be seen in Table 1 below.

Table 1. Tweet Data Samples

index	***Tweet***
0	Ingat! Cantumkan Harta Saat Lapor SPT Tahunan Pajak https://t.co/pUlQMjTkZv
1	@CNNIndonesia Dasar MALING BAJINGAN tengik bisa ngeles juga... Berapa banyak rakyat Indonesia yang hidup miskin, uangnya rakyat untuk foya-foya ASN dirjen pajak, Kemenkeu, beacukai...?
2	@bendvi setuju mas! i've been travelled around the world and i can i say customs indonesia much better drpd negara lain. sepengalamanku jg kl ngenain pajak sesuai dgn aturan kok ini yg heboh ini itu cm krn gaterima aja ngeluarin uang utk penerimaan negara pdhl hrsny dia tau itu konsekuen
3	Haduh..\n\nYg namanya lapor pajak mana ada yg pake 1 pembukuan di Indonesia ini..\n\nLedger 3..\n1. Pemegang saham..\n2. Internal..\n3. Pajak..\n\nBeda2 semua..\n\nLoe lapor semua mana bs cuan gede..\nApalagi pajak..\n\nBayar PPN, PPh badan, PPh Pribadi..\n\nSulap2 sm 'Konsultan' kek Alun..
4	@susipudjiastuti Indonesia dapet cukai import, per kg dpt cukai 1000 perak/ kg. Maka Indonesia akan mendapatkan pajak import 1000x1 000 000 000 =;Rp 1 000 T... mantaaap..
⋮	⋮
7479	@sutanmangara @agushaes01 Setelah pak Prabowo masuk dalam kabinet Pak jokowi..\nData 11ribu trilyun di berikan..\nMakanya Pak Jokowi lewat menteri keuangan mengeluarkan kebijakan Tax amnesty pajak bagi orang2 indonesia yg punya pengahasilan dan hasilnya di taruh di luar negeri..\nPAHAM..drun https://t.co/UaifQGe7l1

Data Preprocessing

Data preprocessing is an essential phase in the preparation of data prior to sentiment extraction and classification. Its primary objective is to enhance the quality of text data. In the context of this study, the data will undergo a comprehensive data preprocessing stage. This stage includes various steps such as Punctuation Removal, Case Folding, Elimination of URLs and usernames, Stopword Removal, Word Normalization, Stemming, and ultimately, Sentiment Labeling. Each of these preprocessing steps plays a crucial role in refining the text data, making it more amenable to subsequent sentiment analysis and classification processes [23].

Dataset Partition

After pre-processing, tweet data undergoes division into two subsets using the HoldOut method, serving as both training and testing data. This split adheres to an 80:20 ratio, with 80% allocated to training data and 20% to test data. The choice of an 80% training and 20% test division is based on strong considerations. Firstly, employing 80% of the data for training enables the model to train on a substantial dataset, facilitating the detection of significant patterns and trends within the dataset. Secondly, dedicating 20% to the test data ensures a reasonably accurate representation of the data used during the modeling process. Finally, this approach enhances computational efficiency by reducing the time required for model training and evaluation. Consequently, this proportion-based division enhances the effectiveness and reliability of the machine learning process [24].

Feature Extraction

Subsequently, the partitioned tweet data proceeds to the feature extraction phase, which employs the TF-IDF (Term Frequency-Inverse Document Frequency) method. The feature extraction aims to identify the most pertinent and informative words within the tweet data. These extracted features will serve as inputs in the classification process. In this context, the collected tweet data is treated as individual documents, while the entire aggregation of tweet data is regarded as a corpus [25]. The formulation of the use of TF-IDF as feature extraction is shown in the equation (1),(2), and (3) below.

$$TF_{ij} = \frac{f_d(i)}{\underset{j \in d}{Max\, f_d(j)}} \tag{1}$$

$$IDF\ (t_i, D) = \log\left(\frac{N}{df(t) + 1}\right) \tag{2}$$

So that,

$$\text{TFIDF} = tf_{d,t} \times idf_t \tag{3}$$

Classification

Sentiment classification in tweet data utilizes the Multinomial Naive Bayes model, which leverages word occurrence probabilities in each tweet relative to their corresponding sentiment class. This model assumes that each word in a tweet is independent of others and equally influences the sentiment class [26]. The process of forming a Naive Bayes classification model can be implemented by calculating the maximum posterior probability value. The calculation of the posterior maximum value is based on the calculation of prior probability and conditional probability (likelihood probability) using training data (training data). In sentiment classification, the calculation of the posterior maximum value can be done using the formula in the equation (4):

$$C_{MAP} = \underset{c \in C}{argmax}\ P(c|d) = \underset{c \in C}{argmax} P(c) \prod_{i=1}^{n} P(w_i|C) \tag{4}$$

Dimana:
CMAP = The highest probabilities for each class
P(c|d) = The conditional probability of class C given D
P(w|c) = Probability of the occurrence of a word (w_i) in a document of class (c)
P(c) = Initial likelihood of a document belonging to class C.

Evaluation

Evaluation refers to the process of assessing the performance of a prediction or classification model using previously unused data [27]. Following sentiment classification using Multinomial Naive Bayes, an evaluation of the model is conducted to assess its performance in classifying sentiment in tweet data. Model evaluation aims to determine the extent to which the created model excels in sentiment classification in tweet data. The model evaluation is carried out by applying several evaluation metrics, including accuracy, precision, and recall (sensitivity). Accuracy measures how well the model correctly classifies data compared to the entire available dataset and can be calculated using the following equation (5).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{5}$$

Precision is intended to assess the extent to which the model accurately classifies data as positive, compared to the total data classified as positive by the model. Precision can be calculated using the following equation (6).

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{6}$$

The purpose of Sensitivity (recall) is to measure how effectively the model correctly classifies data as positive in comparison to the total data that is actually positive. Sensitivity can be calculated using the following equation (7).

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \tag{7}$$

RESULT AND DISCUSSION

A total of 7,480 tweet data were collected and subsequently underwent data preprocessing. This preprocessing phase encompassed various procedures, including Punctuation Removal, Case Folding, URL and Username Elimination, Stopword Removal, Word Normalization, Stemming, and Sentiment Labeling. The primary objective of data preprocessing is to sanitize, transform, and organize the data acquired through web scraping, rendering it suitable and well-structured for data analysis.

The Punctuation Removal process was executed using a Python regex pattern. Notably, an exception was applied to the "@" symbol to facilitate the handling of usernames and mentions in the subsequent URL and username elimination phase. Examples of data resulting from the Punctuation Removal process can be observed in Table 2 below.

Table 2. Example of punctuation removal.

Tweet	*Removing Punctuation*
Ingat! Cantumkan Harta Saat Lapor SPT Tahunan Pajak https://t.co/pUlQMjTkZv	Ingat Cantumkan Harta Saat Lapor SPT Tahunan Pajak httpstcopUlQMjTkZv
@CNNIndonesia Dasar MALING BAJINGAN tengik bisa ngeles juga... Berapa banyak rakyat Indonesia yang hidup miskin, uangnya rakyat untuk foya-foya ASN dirjen pajak, Kemenkeu, beacukai...?	@CNNIndonesia Dasar MALING BAJINGAN tengik bisa ngeles juga Berapa banyak rakyat Indonesia yang hidup miskin uangnya rakyat untuk foyafoya ASN dirjen pajak Kemenkeu beacukai
@bendvi setuju mas! i've been travelled around the world and i can i say customs indonesia much better drpd negara lain. sepengalamanku jg kl ngenain pajak sesuai dgn aturan kok ini yg heboh ini itu cm krn gaterima aja ngeluarin uang utk penerimaan negara pdhl hrsny dia tau itu konsekuen	@bendvi setuju mas ive been travelled around the world and i can i say customs indonesia much better drpd negara lain sepengalamanku jg kl ngenain pajak sesuai dgn aturan kok ini yg heboh ini itu cm krn gaterima aja ngeluarin uang utk penerimaan negara pdhl hrsny dia tau itu konsekuen

Subsequently, the Case Folding process is executed, aiming to convert all characters in the tweet data to lowercase. The case folding process is performed using the str.lower() function in the Python programming language. An example of data that has undergone the case folding process is presented in Table 3 below.

Table 3. Example of case folding

Removing Punctuation	Case Folding
Ingat Cantumkan Harta Saat Lapor SPT Tahunan Pajak httpstcopUlQMjTkZv	ingat cantumkan harta saat lapor spt tahunan pajak httpstcopulqmjtkzv
@CNNIndonesia Dasar MALING BAJINGAN tengik bisa ngeles juga Berapa banyak rakyat Indonesia yang hidup miskin uangnya rakyat untuk foyafoya ASN dirjen pajak Kemenkeu beacukai	@cnnindonesia dasar maling bajingan tengik bisa ngeles juga berapa banyak rakyat indonesia yang hidup miskin uangnya rakyat untuk foyafoya asn dirjen pajak kemenkeu beacukai
@bendvi setuju mas ive been travelled around the world and i can i say customs indonesia much better drpd negara lain sepengalamanku jg kl ngenain pajak sesuai dgn aturan kok ini yg heboh ini itu cm krn gaterima aja ngeluarin uang utk penerimaan negara pdhl hrsny dia tau itu konsekuen	@bendvi setuju mas ive been travelled around the world and i can i say customs indonesia much better drpd negara lain sepengalamanku jg kl ngenain pajak sesuai dgn aturan kok ini yg heboh ini itu cm krn gaterima aja ngeluarin uang utk penerimaan negara pdhl hrsny dia tau itu konsekuen

URL and username elimination is performed to remove Uniform Resource Locators (URLs) and user mentions in each tweet data. The URL and username elimination process is executed using regular expressions in the Python programming language. An example of data that has undergone URL and username removal can be seen in Table 4 below.

Table 4. Example of URL and Username Removal

Case Folding	URL & Username Removal
ingat cantumkan harta saat lapor spt tahunan pajak httpstcopulqmjtkzv	ingat cantumkan harta saat lapor spt tahunan pajak
@cnnindonesia dasar maling bajingan tengik bisa ngeles juga berapa banyak rakyat indonesia yang hidup miskin uangnya rakyat untuk foyafoya asn dirjen pajak kemenkeu beacukai	dasar maling bajingan tengik bisa ngeles juga berapa banyak rakyat indonesia yang hidup miskin uangnya rakyat untuk foyafoya asn dirjen pajak kemenkeu beacukai
@bendvi setuju mas ive been travelled around the world and i can i say customs indonesia much better drpd negara lain sepengalamanku jg kl ngenain pajak sesuai dgn aturan kok ini yg heboh ini itu cm krn gaterima aja ngeluarin uang utk penerimaan negara pdhl hrsny dia tau itu konsekuen	setuju mas ive been travelled around the world and i can i say customs indonesia much better drpd negara lain sepengalamanku jg kl ngenain pajak sesuai dgn aturan kok ini yg heboh ini itu cm krn gaterima aja ngeluarin uang utk penerimaan negara pdhl hrsny dia tau itu konsekuen

Following the elimination of URLs and usernames, the subsequent step involves the normalization process to transform non-standard language words into standardized Indonesian language words. Normalization is performed based on the list of standardized words found in the Colloquial Indonesian Lexicon or Kamus Alay Bahasa Indonesia [28]. During the normalization phase, the text from a tweet is split into separate words using the .split() function in the Python programming language. Following this, a matching process is executed to align these words with the list of standardized words present in the Colloquial Indonesian Lexicon. In cases of correspondence, non-standard words are substituted with the appropriate standardized words. However, if no match is found in the list, the non-standard words remain unchanged. Examples of data that have undergone normalization can be observed in Table 5 below.

Table 5. Example of word normalization

Elimination URL *& Username*	Word Normalization
ingat cantumkan harta saat lapor spt tahunan pajak	ingat cantumkan harta saat lapor spt tahunan pajak
dasar maling bajingan tengik bisa ngeles juga berapa banyak rakyat indonesia yang hidup miskin uangnya rakyat untuk foyafoya asn dirjen pajak kemenkeu beacukai	dasar maling bajingan tengik bisa ngeles juga berapa banyak rakyat indonesia yang hidup miskin uangnya rakyat untuk foyafoya asn dirjen pajak kemenkeu beacukai
setuju mas ive been travelled around the world and i can i say customs indonesia much better drpd negara lain sepengalamanku jg kl ngenain pajak sesuai dgn aturan kok ini yg heboh ini itu cm krn gaterima aja ngeluarin uang utk penerimaan negara pdhl hrsny dia tau itu konsekuen	setuju mas ive been travelled around the world and i can i sayang customs indonesia much better daripada negara lain sepengalamanku juga kalo ngenain pajak sesuai dengan aturan kok ini yang heboh ini itu cuma karena gaterima saja mengeluarkan uang untuk penerimaan negara padahal harusny dia tau itu konsekuen

The next preprocessing step is Stopword Removal. This stage is intended to eliminate non-meaningful words that do not convey essential information in sentiment analysis, such as "and," "or," "in," and the like. Stopword Removal is conducted using Python and the NLTK library in the Indonesian language. Examples of data that have undergone Stopword Removal can be seen in Table 6 below.

Table 6. Example of stopword removal

Word Normalization	*Stopword Removal*
ingat cantumkan harta saat lapor spt tahunan pajak	cantumkan harta lapor spt tahunan pajak
dasar maling bajingan tengik bisa ngeles juga berapa banyak rakyat indonesia yang hidup miskin uangnya rakyat untuk foyafoya asn dirjen pajak kemenkeu beacukai	dasar maling bajingan tengik ngeles rakyat indonesia hidup miskin uangnya rakyat foyafoya asn dirjen pajak kemenkeu beacukai
setuju mas ive been travelled around the world and i can i sayang customs indonesia much better daripada negara lain sepengalamanku juga kalo ngenain pajak sesuai dengan aturan kok ini yang heboh ini itu cuma karena gaterima saja mengeluarkan uang untuk penerimaan negara padahal harusny dia tau itu konsekuen	setuju mas ive been travelled around the world and i can i sayang customs indonesia much better negara sepengalamanku kalo ngenain pajak sesuai aturan heboh gaterima mengeluarkan uang penerimaan negara harusny tau konsekuen

Next, the Stemming process is carried out to transform words into their base form using the Nazief & Adriani Algorithm [29]. The application of the Nazief & Adriani Algorithm is supported by the PySastrawi library in the Python programming language [30]. For example, the stemming results can be seen in Table 7 below.

Table 7. Example of stemming

Stopword Removal	*Stemming*
cantumkan harta lapor spt tahunan pajak	cantum harta lapor spt tahun pajak
dasar maling bajingan tengik ngeles rakyat indonesia hidup miskin uangnya rakyat foyafoya asn dirjen pajak kemenkeu beacukai	dasar maling bajing tengik ngeles rakyat indonesia hidup miskin uang rakyat foyafoya asn dirjen pajak kemenkeu beacukai
setuju mas ive been travelled around the world and i can i sayang customs indonesia much better negara sepengalamanku kalo ngenain pajak sesuai aturan heboh gaterima mengeluarkan uang penerimaan negara harusny tau konsekuen	tuju mas ive been travelled around the world and i can i sayang customs indonesia much better negara alam kalo ngenain pajak sesuai atur heboh gaterima keluar uang terima negara harusny tau konsekuen

The final step in data preprocessing is sentiment labeling. The labeling process aims to assign sentiment labels to tweet data based on the polarity values or the level of emotional tendency within each tweet. Polarity values are determined using the InSet (Indonesian Sentiment Lexicon) dictionary, which ranges from -5 (very negative) to +5 (very positive) [31].

During the labeling stage, each word in a tweet is tokenized using the .split() function in Python. Subsequently, each token is matched with the list of words and their corresponding polarity values in InSet. As a result, the polarity values of each word are aggregated to determine the emotional tendency within the tweet. The formulation for this labeling process is described in Equation (8) below.

$$\begin{aligned} Sentimen\ Class &= t_1 + t_2 + t_3 + t_4 + t_5 + ..\, t_n \\ &= total\ sentiment\ polarity \end{aligned} \tag{8}$$

If the sentiment score exceeds 0, the tweet will be classified as "Positive." Conversely, if the sentiment score is less than 0, the tweet will be categorized as "Negative." Meanwhile, if the sentiment score is equal to 0, the tweet will be declared as "Neutral." You can see an example of the labeling results in Table 8 below

Table 8. Example of labeling

Tweet	Polarity Value	Tweet Polarity Sum	Sentiment label
cantum harta lapor spt tahun pajak	[('cantum', 4), ('lapor', 2), ('pajak', -3)]	3	*Positive*

dasar maling bajing tengik ngeles rakyat indonesia hidup miskin uang rakyat foyafoya asn dirjen pajak kemenkeu beacukai	[('bajing', -5), ('ngeles', -3), ('hidup', -4), ('miskin', -5), ('pajak', -3)]	-20	*Negative*
tuju mas ive been travelled around the world and i can i sayang customs indonesia much better negara alam kalo ngenain pajak sesuai atur heboh gaterima keluar uang terima negara harusny tau konsekuen	[('tuju', -4), ('sayang', -3), ('better', -1), ('alam', -1), ('pajak', -3), ('sesuai', 3), ('atur', -4), ('heboh', -3), ('keluar', -3), ('terima', 2), ('tau', -4)]	-21	*Negative*

Based on the labeling results, there were 572 tweets showing a positive sentiment, while 6706 tweets expressed a negative sentiment, and 202 tweets depicted a neutral sentiment. The ratio of the total for these sentiment classes can be observed in Figure 2 below.

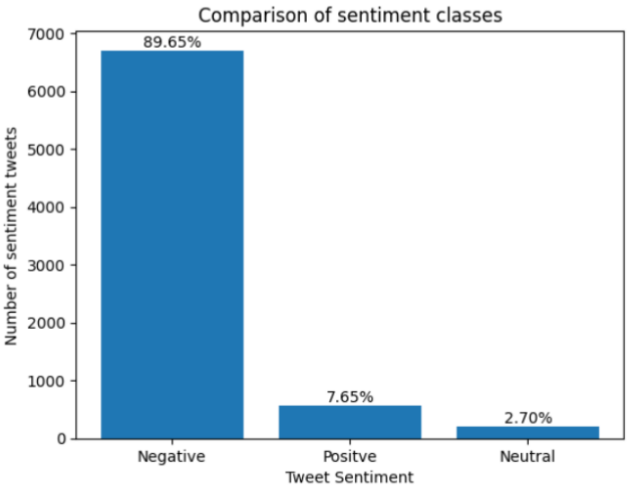

Figure 2. Comparison of sentiment classes

The comparison between the number of tweets expressing negative sentiment significantly surpasses the count of tweets expressing positive and neutral sentiments. This comparison results in an indication that approximately 89.65% of tweets related to taxation topics in Indonesia tend to carry a negative emotional tone. Upon completion of labeling, the subsequent step involves dividing the tweet data using the holdout method before extraction using TF-IDF. The data splitting scenario consists of 80% training data and 20% test data, with a total of 5,984 training data and 1,496 test data utilized. The Frequency-Inverse Term Document Frequency (TF-IDF) method is employed for feature extraction. This extraction refers to equations (1), (2), and (3). The TF-IDF weighting scheme is outlined in Tables 9 and 10 below.

Table 9.Sample data for sentiment weighting scenario.

Data index	*Tweet*
15	tanya dong apa jadi kalo semua rakyat indonesia enggak bayar pajak
16	kok bisa pajak rakyat indonesia buat bayar bapak satu ini
4381	bayar pajak adalah bentuk komitmen dalam bangun indonesia

Table 10. TF-IDF Weighting Scenario Using Sample Data

Term (t)	TF $TF_{ij} = \frac{f_d(i)}{Max\ f_d(j)}$ $j \in d$			DF	IDF	$tf_{d,t} \times idf_t$		
	15	16	4381		Log$\frac{n}{DF+1}$	15	16	4381
Tanya	$\frac{1}{11} = 0{,}090$	0	0	1	1.098	0.099	0	0
Pajak	$\frac{1}{11} = 0{,}090$	$\frac{1}{10} = 0{,}100$	$\frac{1}{8} = 0{,}125$	3	0	0	0	0

Indonesia	$\frac{1}{11} = 0{,}090$	$\frac{1}{10} = 0{,}100$	$\frac{1}{8} = 0{,}125$	3	0	0	0	0
Bayar	$\frac{1}{11} = 0{,}090$	$\frac{1}{10} = 0{,}100$	$\frac{1}{8} = 0{,}125$	3	0	0	0	0
Rakyat	$\frac{1}{11} = 0{,}090$	$\frac{1}{10} = 0{,}100$	0	2	0	0	0	0
Komitmen	0	0	$\frac{1}{8} = 0{,}125$	1	1.098	0	0	0.137
Bangun	0	0	$\frac{1}{8} = 0{,}125$	1	1.098	0	0	0.137

Based on the calculations presented in Table 10 above, the word feature "tanya" exhibits a TF-IDF value of 0.099 at the data index 15, while it is entirely absent at the data indices 16 and 4381, with a TF-IDF value of 0. This result indicates that the word "tanya" is not found at indices 16 and 4381 due to having a TF-IDF value of 0.With the assistance of the TfidfVectorizer() function in Python, a total of 11,719 word features were discovered within the dataset of 7,480 tweet data. Some of these word features can be found in Table 11 below.

Table 11.Sample of TF-IDF word features in tweet data

Feature index	TF-IDF Feature
0	Aaah
1	Aab
10	Abang
11	Abangda
944	Bayarr
966	Beacukai
1036	Belikan
· ⋮ ⋮	⋮ ⋮
11718	zudan

Subsequently, sentiment classification in tweet data is conducted using the Multinomial Naive Bayes model, wherein the conditional probabilities are computed based on the frequency of appearance of each term (t) within each sentiment class (C). The calculation of these conditional probability values is predicated on TF-IDF weighting. The formulation of the Multinomial Naive Bayes model along with TF-IDF weighting is presented in Equation (9), as referenced in Equation (4).

$$P(C|d) \propto P(C) \times \prod_{t \in d} P(t\,|C) \tag{9}$$

The value of $P(C)$ can be computed using the formula in Equation (10).

$$\mathrm{P(c)} = \frac{N_c}{N} \tag{10}$$

Then, The value of $P(t\,|C)$ is calculated using the formula in Equation (11).

$$P(t|C) = \frac{w_{c,t} + 1}{(\sum_{w' \in V} w_{c,t'}) + B'} \tag{11}$$

The classification process for the entire test data is executed using the Python programming language. The data, which has been extracted using the TfidfVectorizer() function, will be classified using the MultinomialNB() model.

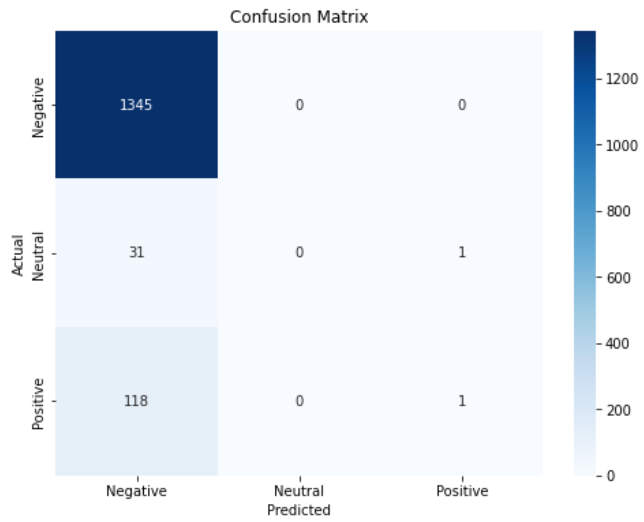

Figure 3. Classification Confusion Matrix

Visualization of the Confusion Matrix (Figure 3) reveals that 1345 actual data with negative sentiment were correctly predicted in the negative class, 31 actual data with neutral sentiment were incorrectly predicted as negative class, 1 actual data with neutral sentiment was erroneously predicted as positive class, 118 actual data with positive sentiment were misclassified as negative class, and 1 actual data with positive sentiment were accurately predicted in the positive class. These prediction outcomes reflect the classification model's tendency to identify test data as negative sentiment. This tendency is due to the dominance of training data with negative sentiment, which significantly outnumber positive and neutral sentiment in the training dataset. Imbalance in this dataset has the potential to impact the model's performance in making predictions.

The step to address dataset imbalance involves random undersampling of the negative sentiment data (the Majority Data) [32]. The purpose of the random undersampling process is to reduce the number of samples in the majority class to equal that of the minority class, thereby creating a balanced distribution among positive, neutral, and negative sentiments. The random undersampling process is carried out through the following steps [33]:

1. Identify the majority class (the category with the most samples) and the minority class (the category with the fewest samples).
2. Determine the number of samples to be retained from the majority class. This number can be determined based on the desired proportion between the majority and minority classes, such as using a specific ratio or an absolute count..
3. Randomly select the specified number of samples from the majority class to retain..
4. Combine the retained samples from the majority class with the samples from the minority class to form a balanced dataset.

After random undersampling, the total dataset comprises 606 instances, divided evenly with 202 samples each for positive, neutral, and negative sentiments. The classification results from this random undersampled data, as depicted in Figure 4, indicate that 28 instances were correctly classified as negative. Among these, 11 negative instances were misclassified as neutral, and 6 were incorrectly classified as positive. Additionally, 9 neutral instances were misclassified as negative, 14 were accurately predicted as neutral, and 14 were wrongly categorized as positive. Furthermore, 5 positive instances were misclassified as negative, 12 were misclassified as neutral, and 23 were correctly predicted as positive. The utilization of the random undersampling method facilitates a balanced distribution of data across sentiment categories, enabling the model to yield more balanced prediction outcomes within the context of this research.Proses

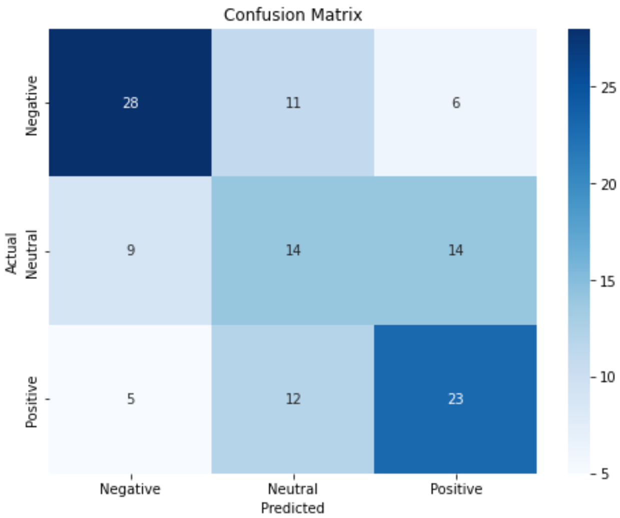

Figure 4. Confusion matrix for random undersampling data

The model evaluation was conducted under two testing scenarios: classification on data before random undersampling and after undergoing random undersampling. Both scenarios were utilized to analyze and compare the model's performance in classifying data under different conditions. The comparative results of these two scenarios are presented in Table 12 below.

Table 12. Classification Result Comparison

Before	*Confusion Matrix*		*Predicted*			*Accuracy*	*Recall*	*Precision*
			Negative	Neutral	Positive			
	Actual	Negative	1345	0	0	89.97%	33.61%	46.68%
		Neutral	31	0	1			
		Positive	118	0	1			
After	*Confusion Matrix*		*Predicted*			*Accuracy*	*Recall*	*Precision*
			Negative	Neutral	Positive			
	Actual	Negative	28	11	6	53.28%	52.52%	52.66%
		Neutral	9	14	14			
		Positive	5	12	23			

CONCLUSION

Based on the research findings and analysis conducted, it was discovered that out of a total of 7480 analyzed tweet data, 572 tweets exhibited positive sentiment, 6706 tweets were negative, and 202 tweets were neutral. This indicates that approximately 89.65% of tweets related to taxation in Indonesia tend to have a negative sentiment. Furthermore, the study also revealed that the TF-IDF weighting method and Naive Bayes classification can be employed to analyze taxation sentiment with an accuracy rate of 89.97%, precision of 46.68%, and sensitivity of 33.61%. However, the classification model tended to predict test data as negative sentiment class. Therefore, a random undersampling technique was applied to achieve a more balanced distribution between positive, neutral, and negative sentiment classes. The results of the random undersampling technique yielded an accuracy of 53.28%, precision of 52.66%, and sensitivity of 52.52%. Nevertheless, the model still faced challenges in classifying data with positive and neutral sentiments, primarily due to the dominance of tweet data with negative sentiment.

REFERENCES

[1] Republik Indonesia, “Undang-Undang Republik Indonesia Nomor 9 Tahun 1994 Tentang Perubahan Atas Undang-Undang Nomor 6 Tahun 1983 Tentang Ketentuan Umum Dan Tata Cara Perpajakan.” 1994. [Online]. Available: Https://Www.Bphn.Go.Id/Data/Documents/94uu009.Pdf

[2] D. R. Niru And A. Sinaga, “Pemungutan Pajak Dan Permasalahannya Di Indonesia,” Vol. 7, No. 1, 2016.
[3] M. Ismayadie, “Analisis Efektivitas Dan Kontribusi Penerimaan Pajak Bumi Dan Bangunan (PBB) Dan Pajak Penghasilan (PPH) Terhadap Pendapatan Negara Tahun 2007-2017,” *Equity: Jurnal Ekonomi*, Vol. 7, No. 2, Pp. 12–24, Dec. 2019, Doi: 10.33019/Equity.V7i2.5.
[4] I. R. S. Rahayu, “Sepanjang 2022, Penerimaan Pajak Tembus Rp 1.716 Triliun,” *Kompas*, Jan. 03, 2023.
[5] Ç. Erhan And S. \.Isa, “A Brief Overview Of The Tax Compliance Literature Specific To Experimental Studies,” *Journal Of Public Finance Studies*, No. 69, Pp. 187–210, 2023.
[6] Z. Ibrahim, A. Ibrahim, And Syahribulan, “Pengaruh Kepercayaan Publik Terhadap Kepatuhan Wajib Pajak,” *YUME : Journal Of Management*, Vol. 3, No. 3, Pp. 80–93, 2020.
[7] H. Azizah, B. S. Rintyarna, And T. A. Cahyanto, “Sentimen Analisis Untuk Mengukur Kepercayaan Masyarakat Terhadap Pengadaan Vaksin Covid-19 Berbasis Bernoulli Naive Bayes,” *BIOS: Jurnal Teknologi Informasi Dan Rekayasa Komputer*, Vol. 3, No. 1, Pp. 23–29, 2022.
[8] D. R. Alghifari, M. Edi, And L. Firmansyah, “Implementasi Bidirectional LSTM Untuk Analisis Sentimen Terhadap Layanan Grab Indonesia,” *Jurnal Manajemen Informatika (JAMIKA)*, Vol. 12, No. 2, Pp. 89–99, 2022.
[9] I. Verawati And B. S. Audit, “Algoritma Naïve Bayes Classifier Untuk Analisis Sentiment Pengguna Twitter Terhadap Provider By.U,” *JURNAL MEDIA INFORMATIKA BUDIDARMA*, Vol. 6, No. 3, P. 1411, Jul. 2022, Doi: 10.30865/Mib.V6i3.4132.
[10] B. R. Atmadja, “Analisis Sentimen Bahasa Indonesia Pada Tempat Wisata Di Kabupaten Sukabumi Dengan Naïve Bayes,” Vol. 15, No. 2, Pp. 371–382, 2022.
[11] C. Andrade, “The Limitations Of Online Surveys,” *Indian J Psychol Med*, Vol. 42, No. 6, Pp. 575–576, 2020.
[12] J. P. Agans, S. A. Schade, S. R. Hanna, S.-C. Chiang, K. Shirzad, And S. Bai, “The Inaccuracy Of Data From Online Surveys: A Cautionary Analysis,” *Qual Quant*, Pp. 1–22, 2023.
[13] M. K. Durán-Vaca And J. A. Ballesteros-Ricaurte, “Sentiment Analysis On Twitter To Measure The Perception Of Taxation In Colombia,” In *International Conference Europe Middle East & North Africa Information Systems And Technologies To Support Learning*, 2019, Pp. 184–193.
[14] J. P. Gujjar And H. P. Kumar, “Sentiment Analysis: Textblob For Decision Making,” *Int. J. Sci. Res. Eng. Trends*, Vol. 7, No. 2, Pp. 1097–1099, 2021.
[15] N. P. Challa, K. R. Madhavi, B. Naseeba, B. B. Bhanu, And C. Naresh, “Sentiment Analysis From TWITTER Using NLTK,” In *International Conference On Hybrid Intelligent Systems*, 2023, Pp. 852–861.
[16] A. S. R. Rufaida, A. E. Permanasari, And N. A. Setiawan, “Lexicon-Based Sentiment Analysis Using Inset Dictionary: A Systematic Literature Review,” In *ICAE 2022: Proceedings Of The 5th International Conference On Applied Engineering, ICAE 2022, 5 October 2022, Batam, Indonesia*, 2023, P. 258.
[17] F. Muharram And K. Saputra, “Analisis Sentimen Pengguna Twitter Terhadap Kinerja Walikota Medan Menggunakan Metode Naive Bayes Classifier,” *Jurnal Sistem Informasi Dan Ilmu Komputer*, Vol. 1, No. 2, Pp. 1–12, 2023.
[18] A. I. Kadhim, “Term Weighting For Feature Extraction On Twitter: A Comparison Between BM25 And TF-IDF,” In *2019 International Conference On Advanced Science And Engineering (ICOASE)*, 2019, Pp. 124–128.
[19] O. S. D. Silaen, H. Herlawati, And R. Rasim, “Analisis Sentimen Mengenai Gangguan Bipolar Pada Twitter Menggunakan Algoritma Naïve Bayes,” *Jurnal Komtika (Komputasi Dan Informatika)*, Vol. 6, No. 2, Pp. 62–73, Nov. 2022.
[20] A. Miftahusalam, A. Febby Nuraini, A. A. Khoirunisa, And H. Pratiwi, “Perbandingan Algoritma Random Forest, Naïve Bayes, Dan Support Vector Machine Pada Analisis Sentimen Twitter

Mengenai Opini Masyarakat Terhadap Penghapusan Tenaga Honorer," In *SEMINAR NASIONAL OFFICIAL STATISTICS 2022* , 2022, P. 563.

[21] D. Farah Zhafira, B. Rahayudi, And P. Korespondensi, "ANALISIS SENTIMEN KEBIJAKAN KAMPUS MERDEKA MENGGUNAKAN NAIVE BAYES DAN PEMBOBOTAN TF-IDF BERDASARKAN KOMENTAR PADA YOUTUBE," Vol. 2, No. 1, Pp. 55–63, 2021.

[22] H. M. Saragih, "Analisis Sentimen Pengguna Twitter Terhadap Layanan Pajak Kendaraan Bermotor Menggunakan Algoritme Naive Bayes Classifier," 2021.

[23] H.-T. Duong And T.-A. Nguyen-Thi, "A Review: Preprocessing Techniques And Data Augmentation For Sentiment Analysis," *Comput Soc Netw*, Vol. 8, No. 1, Pp. 1–16, 2021.

[24] D. S. Abdelminaam, N. Neggaz, I. A. E. Gomaa, F. H. Ismail, And A. A. Elsawy, "Arabicdialects: An Efficient Framework For Arabic Dialects Opinion Mining On Twitter Using Optimized Deep Neural Networks," *Ieee Access*, Vol. 9, Pp. 97079–97099, 2021.

[25] L. Chiruzzo, P. Amarilla, A. R\'\Ios, And G. Giménez-Lugo, "Development Of A Guarani-Spanish Parallel Corpus," In *Proceedings Of The Twelfth Language Resources And Evaluation Conference*, 2020, Pp. 2629–2633.

[26] E. Alpaydin, *Introduction To Machine Learning*, 3rd Ed. Cambridge: The MIT Press, 2014.

[27] L. Zhu, J. Li, Y. La, And T. Jia, "Improving The Accuracy Of Remote Sensing Land Cover Classification By GEO-ECO Zoning Coupled With Geostatistical Simulation," *Applied Sciences*, Vol. 11, No. 2, P. 553, 2021.

[28] N. A. Salsabila, Y. A. Winatmoko, A. A. Septiandri, And A. Jamal, "Colloquial Indonesian Lexicon," In *International Conference On Asian Language Processing (IALP)*, Bandung: IEEE, Jan. 2019.

[29] S. F. S. Sodiq, W. Desena, A. Wibowo, And Others, "Penerapan Algoritma Stemming Nazief & Adriani Pada Proses Klasterisasi Berita Berdasarkan Tematik Pada Laman (Web) Direktorat Jenderal HAM Menggunakan Rapidminer," *Syntax: Jurnal Informatika*, Vol. 11, No. 02, Pp. 10–21, 2022.

[30] A. A. Hidayat And B. Pardamean, "Count Time Series Modelling Of Twitter Data Topic Modelling: A Case Of Indonesia Flood Events," In *AIP Conference Proceedings*, 2023.

[31] F. Koto And G. Y. Rahmaningtyas, "Inset Lexicon: Evaluation Of A Word List For Indonesian Sentiment Analysis In Microblogs," In *Proceedings Of The 2017 International Conference On Asian Language Processing, IALP 2017*, Institute Of Electrical And Electronics Engineers Inc., Feb. 2018, Pp. 391–394. Doi: 10.1109/IALP.2017.8300625.

[32] B. Liu And G. Tsoumakas, "Dealing With Class Imbalance In Classifier Chains Via Random Undersampling," *Knowl Based Syst*, Vol. 192, P. 105292, 2020.

[33] Z. Wang, C. Cao, And Y. Zhu, "Entropy And Confidence-Based Undersampling Boosting Random Forests For Imbalanced Problems," *IEEE Trans Neural Netw Learn Syst*, Vol. 31, No. 12, Pp. 5178–5191, 2020.