JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, Agustus 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8922
Hal 555-566
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Analysis of Public Sentiment Towards Tax Increases Impacting Unemployment Using SVM and Multinomial Naive Bayes Methods

Siti Nurhaliza Sofyan*, Zulham Sitorus, Muhammad Iqbal

Postgraduate, Master of Information Technology, Panca Budi University of Development , Medan, Indonesia
Email: [1*]sitinurhalizas102@gmail.com, [2]zulhamsitorus@dosen.pancabudi.ac.id, [3]muhammadiqbal@dosen.pancabudi.ac.id
Email Penulis Korespondensi : sitinurhalizas102@gmail.com
Submitted **23-07-2025;** Accepted **14-08-2025;** Published **16-08-2025**

Abstract

Tax increase policies often generate pros and cons among the public, especially when perceived as having an impact on increasing unemployment. This study aims to analyze public sentiment regarding the issue of tax increases impacting unemployment by utilizing Machine Learning classification methods, namely Support Vector Machine (SVM) and Multinomial Naive Bayes (MNB). The data used comes from social media platform X in the form of public opinions collected online and then categorized into three sentiments: positive, negative, and neutral, with a total of 1,000 sentiment data points. The analysis process included text preprocessing, feature extraction with TF-IDF, and classification using both methods. In the Test and Score algorithm, the SVM algorithm produced an AUC of 0.660, CA of 0.694, F1 of 0.569, and Recall of 0.694, while the MNB algorithm produced an AUC of 0.586, CA of 0.198, F1 of 0.105, and Recall of 0.198. The study concluded that Support Vector Machines (SVMs) had a higher level of accuracy than Multinominal Naïve Bayes in classifying public sentiment. The majority of public opinion tended to be negative, indicating concern about the impact of tax increases on the workforce. These findings provide important insights for policymakers to consider public perception when establishing future fiscal policy.

Keywords: Sentiment Analysis, Tax Increase, Unemployment, SVM, Multinomial Naive Bayes

1. INTRODUCTION

Taxes are one of the main sources of state revenue and play a strategic role in supporting national development. In the context of fiscal policy, the government often adjusts tax rates to increase state revenue and maintain economic stability. However, tax increase policies are not always accepted positively by all segments of society. On the contrary, such policies often trigger negative reactions, especially when perceived as burdensome for the public or the business sector. One of the concerns arising from tax hikes is the potential increase in unemployment rates, which stems from economic pressures on businesses, leading to reduced production capacity or workforce reductions [1][2].

With the development of information and communication technology, people are now more active in voicing their opinions and attitudes towards public issues through social media. Platforms such as X (formerly Twitter), Facebook, and other online discussion forums have become new public spaces that reflect social dynamics in real time. Opinions expressed through social media reflect the perceptions, expectations, and criticisms of the public towards government policies, including fiscal policies such as tax increases. Therefore, it is important to capture and analyze the sentiments emerging in society to understand to what extent such policies are accepted or rejected by the public.

In this context, data mining becomes an important process that applies intelligent methods to extract patterns from data [3][4][5][6]. One branch of data mining, namely sentiment analysis, is part of Natural Language Processing (NLP) which aims to recognize and classify emotions or opinions in text, such as user opinions on social media. Through sentiment analysis, in-depth insights can be gained regarding public responses to issues such as tax increases linked to rising unemployment. [7] [8]–[10].

In this study, two methods were used: Support Vector Machine (SVM) and Multinomial Naive Bayes (MNB). SVM is known to be effective in handling high-dimensional data such as text, with the ability to separate data classes using optimal hyperplanes [2][3][13][14][15][16].

Meanwhile, Multinomial Naive Bayes is known as a simple yet effective method, especially in classifying documents or texts based on word frequency. In the journal "The Use of the Naïve Bayes Method in Classifying Unemployment in Bojong Kulur Village," the results obtained using the Naïve Bayes method were quite high, with an accuracy of 80%, precision of 100%, and recall of 50%. In the journal [2] With the title "Application of the K-Means Method in Classifying Unemployment in Indonesia," it was concluded that there are 13 provinces, namely cluster 1 with the highest unemployment rate and cluster 2 with 21 provinces with low unemployment potential. [17].

Based on the current issues, I chose to conduct this research considering that this is the situation we are facing today. With the aim of this research, I seek to analyze the opinions of the public found on social media platforms (X) through the Support Vector Machine (SVM) and Multinomial Naive Bayes (MultinomialNB) methods. Through this analysis, it is hoped that the government or relevant authorities can identify solutions to the issue of tax increases that have led to rising unemployment. It is further hoped that unemployment can be addressed efficiently, enabling the public to secure sustainable employment to meet their daily needs.

2. RESEARCH METHODOLOGY

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, Agustus 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8922
Hal 555-566
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

2.1 Research Diagram

To provide a more systematic overview of the stages involved in this study, the following flowchart is presented. This diagram summarizes the main steps taken from data collection to the final analysis using the Support Vector Machine (SVM) and Multinomial Naive Bayes (MNB) methods. Each stage is designed in a structured manner to ensure that the sentiment analysis of public opinion regarding tax increases that impact unemployment can be conducted effectively and accurately. The research diagram is as follows:

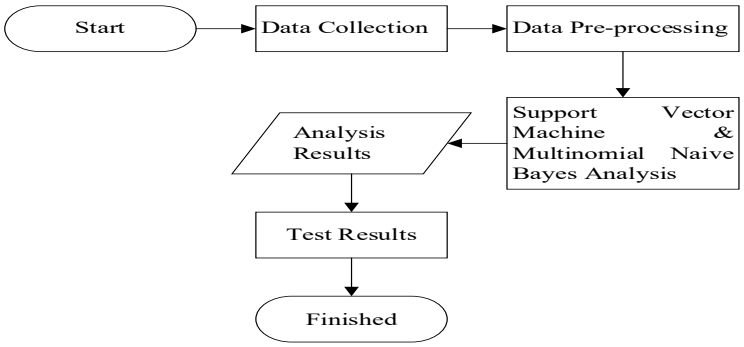

Figure 1. Research Diagram

Based on the diagram above, the research steps can be explained as follows:

1. The research begins with the initiation process, which includes planning the research steps to be taken.
2. Data Collection: The data collection process is carried out to gather the necessary information. This involves collecting comments and opinions from platform X regarding the impact of tax increases on unemployment.
3. Data Preprocessing: After the data is collected, preprocessing is performed to focus on KDD. The cleaning process includes removing duplicate data, checking for inconsistent data, and correcting errors in the data.
4. Support Vector Machine and Multinominal Naive Bayes Analysis: The analysis process uses the Support Vector Machine and Multinominal Naive Bayes algorithms to classify data and create predictive models as well as identify patterns or groups that may exist in the dataset.
5. Analysis Results: A comparison of each algorithm analysis is carried out to determine which method is more accurate and best suited to the research.
6. Testing Results: The analyzed models are then tested using test data to assess the model's generalization ability on new data.
7. Completion: The research is concluded after re-evaluation and ensuring that all research steps have been properly executed. The conclusions and findings of the research can be presented or published. [18]

2.2 Research Stages

Research method [10] The method used is an experimental method by observing the variables of the object being studied. The experimental method aims to test the effect of one variable on another or to test the cause-and-effect relationship between variables. The steps in designing the research are as follows:

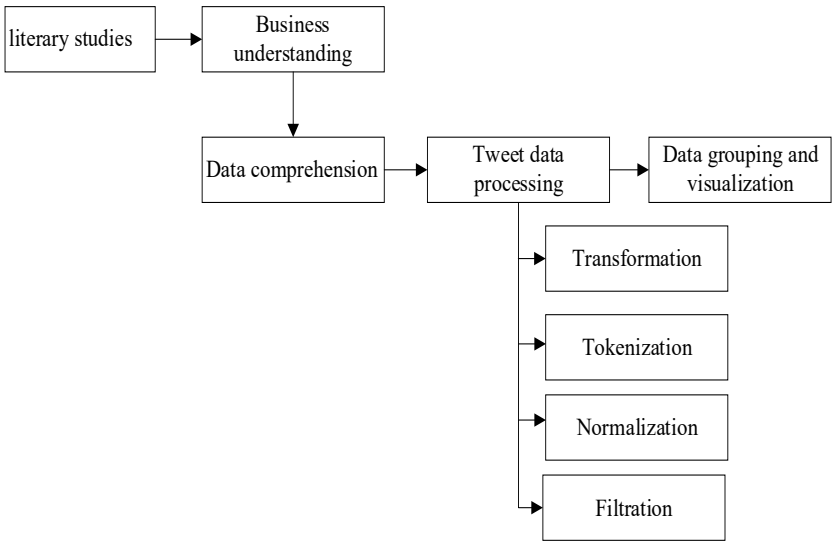

Figure 2. Research Process Flow

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, Agustus 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8922
Hal 555-566
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

2.3 Support Vector Machine (SVM)

2.3.1 Data Preparation

Data preparation stage [11] This is the stage of preparing data before it is used for modeling and evaluation. This stage is divided into two parts, namely:

1. Cleaning.
   During the cleaning process, two things will be done. First, attribute selection will be carried out by deleting several unnecessary attributes such as ID, date, and account name. Second, data will be cleaned of duplicates, empty data, and outliers.
2. Transformation.
   The transformation stage is basically carried out to convert data into a form that can be accepted by the algorithm, because SVM can only accept numerical values, while the data is text data, so changes need to be made. The steps are as follows:
   a. Case folding.
      Case folding aims to convert all documents in a dataset into a similar format, namely lowercase letters. In addition to converting letters to lowercase, this stage also involves removing all unnecessary symbols.
   b. Slang word removal.
      Slang word removal aims to change all words that do not use standard language and replace them with their actual forms.
   c. Stop word removal.
      Stop word removal aims to remove all words that have no meaning, such as conjunctions.
   d. Stemming.
      Stemming aims to change word fragments into their original form by removing prefixes, suffixes, and other word insertions.
   e. Text vectorization.
      Text vectorization is a method of weighting words. At this stage, text data is converted into numerical form so that it can be processed by algorithms. One of the algorithms that is quite often used in text vectorization is term frequency inverse document frequency (IF-IDF).

2.3.2 Modelling

After the data goes through the data preparation stage, it continues with the modeling or training process of the SVM algorithm. At this stage, the model will learn word patterns for defamatory sentences so that it can later group these sentences or tweets. The stages of the training process of the SVM algorithm can be described as follows:

1. Finding the support vector value.
   As mentioned earlier, support vector values play a role in the formation of hyperplanes. One way to find the value of support vectors is by using kernel functions. Kernel functions are functions used to map data, whether it is linearly or non-linearly distributed. There are several kernels that can be used for text classification, but the most commonly used and best performing for text classification is the linear kernel. As the name suggests, this kernel maps data linearly. The reason why this kernel is very good for text classification is, first, because text data usually has a linear distribution, and second, the linear kernel performs well when handling data with a large number of attributes. The formula for the linear kernel is:

$$\chi(i,j) = x_i x_j^{\tau}$$

Where:

x : document or text
i, j : document sequence
T : transpose

2. Finding the alpha value.
   After the support vector value is found, the next step is to find the alpha value, which plays a role in determining the weight value. To find the alpha value, we can use a formula called the Lagrange multiplier, as follows:

$$\max_{\alpha} L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{i,j=1}^{n} \alpha_i \, \alpha_j \; y_i \, y_j \, (x_i^T \, . \, x_j)$$

Where :

a : alpha value y
y : support vector value

3. Find the weight values.
   Once the alpha value has been found, continue by finding the weight values. These weight values determine the slope of the hyperplane. The following formula can be used to find the weight values:

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, Agustus 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8922
Hal 555-566
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

$$\frac{\partial}{\partial w} L_p\,(w, b, \alpha) = 0 \rightarrow w = \sum_{i=1}^{n} \alpha_i\, y_i\, x_i$$

where :

w : weight value

4. Find the bias value.

After the weight values are found, the next step is to find the bias value. The bias value is the value that determines the position of the hyperplane, and the formula for finding the bias value is as follows:

$$b = -\frac{1}{2}\,(w.x^{+} + w.x^{-})$$

Wehere :

b : bias value

5. Determining the class of the text.

After all values have been obtained and the hyperplane has been formed, the next step is to determine the class of the text. The SVM method for determining the class of data is to use a decision function with the following formula:

$$f(x_d) = sign(w.k((x_d)) + b$$

Every value obtained from a decision with a value less than zero will be categorized as negative, or in this case, not defamation, and values greater than zero will be categorized as positive, or defamation.

2.3.3 Model Evaluation Stages

The model is evaluated using accuracy, precision, and recall metrics to ensure its reliability and validity in predicting the risk of dropping out of college. The model evaluation formula is as follows:

$$Akurasi = \frac{TP - TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Precision = \frac{TP}{TP + FN} \quad (3)$$

Where:
TP: True Positive (number of positive data correctly classified)
TN: True Negative (number of negative data correctly classified)
FP: False Positive (number of negative data incorrectly classified as positive)
FN: False Negative (number of positive data incorrectly classified as negative)

2.4 Multinominal Naive Bayes

Multinomial Naive Bayes is a widely used classification algorithm for text data. This algorithm applies a probabilistic method to predict the category of text documents based on word frequency. Bayes' theorem provides a way to update probability estimates for a hypothesis as more evidence or information becomes available.

The general formula of Bayes' theorem, which forms the basis of Naive Bayes:

$$\boldsymbol{P(C|X) = \frac{P(C|X).P(C)}{P(X)}} \quad (4)$$

Where:

X = Sample data with unknown class (label).

C = Hypothesis that X is class (label) data.

P(C) = Probability of hypothesis C.

P(X) = Probability of observed sample data (probability of C).

P(X|C) = Probability based on conditions in the hypothesis.

To calculate the probability of a message falling into a particular category, Multinomial Naive Bayes uses a multinomial distribution.:

$$P\,(\,X\,) = \frac{N!}{N_1!\,N_2!\ldots N_M!}\,P_1^{N_1}\,P_2^{N_2}\ldots P_M^{N_M} \quad (5)$$

Where:

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, Agustus 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8922
Hal 555-566
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

n is the total number of trials.n i adalah jumlah kejadian untuk hasil i.

p i is the probability of outcome i.

To estimate how likely each word is to belong to a particular class such as “spam” or “not spam,” we use a method called Maximum Likelihood Estimation (MLE). This helps find probabilities based on the actual amount of our data. The formula is:

$$\theta_{c,i} = \frac{\text{count}_{(w_i,c)+1}}{N+v} \quad (6)$$

Where:

count ($_{wi,c}$) is the number of times the word wi appears in document class c.

N is the total number of words in document class c c .

$_{v}$ is the size of the vocabulary.

Multinomial Naïve Bayes merupakan metode supervised learning, sehingga setiap data perlu diberikan label sebelum dilakukan training. Probabilitas suatu dokumen *d* berada di kelas *c* dapat dihitung menggunakan Persamaan 1 :

$$P(c|d) \propto P(c) \prod_{k=1}^{n} P(t_k|c) \quad (7)$$

$P(c|d)$: The probability of document d being in class c

$P(c)$: The prior probability of a document is in class c

$\{t1,t1,t1, \dots ,tn\}$:Tokens in document d that are part of a vocabulary with the sum of n

$P(tk|c)$: The conditional probability of the term kindergarten is in the document in class c

The best class in the Naïve Bayes classification is determined by looking for the *maximum a posteriori* (MAP) of the cmap class through equation 2

$$c_{\text{map}} = \arg\max_{c \in C} \hat{P}(c) \prod_{k=1}^{n} \hat{P}(t_k|c)$$

2.5 Research Data

Table 1. Research Data

No	Comments	Sentiment
1	Layoffs where job opportunities are scarce, Yes, one of the factors causing the return migration to be less festive than previous years	negative
2	Tens of thousands of layoffs, weak purchasing power, Sri Mulyani still talks about the economy, Eid al-Fitr celebrations subdued.	negative
3	In my area, people have been returning home since last year because there are no jobs left in the city due to layoffs and bankruptcies.	negative
4	It's not as simple as just removing the rules and everything becoming ideal. That can only be done if the government is able to provide many decent replacement jobs. If it's done immediately, there will be layoffs everywhere, many MSMEs will go bankrupt, and ultimately the country's economy will collapse.	negative
5	The government ensures there will be no layoffs of employees at state-owned palm oil plantations.	positive
6	The government is taking swift action to address layoff issues	positive
7	Please don't raise taxes by ~~12%~~ yet	positive
8	Let's buy more so we can pay the tax #12%tax	neutral
9	Prove it first, Mr. President. Once you have succeeded, then you can say things like this. The professor conducted an analysis primarily so that you would not force the country into more debt by raising taxes by 12%. During the 97 years since Indonesia's independence, we, the people, have still been able to eat	positive
10	Besides increasing taxes to 12%, the taxable items are also increasing. #Thank you, Mr. Prabowo, for the free nutritious lunch. We’re waiting for it—don’t let it just become a new corruption scheme. #12%Tax	negative
11	He said it would take 7 hours, but Arie said it would only take 2 hours, right? The issue has been shifted from the movement to reject the 12% tax increase. Stay alert!	Negative
12	Yes, we believe the public will feel the hardships of the 12% tax.	negative

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, Agustus 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8922
Hal 555-566
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

13	This country is a joke #taxincrease #12percenttax #naturalresources #tax #layoffs #government #people #politics	negative
14	Already hit by the 12% tax hike, the school canteen is being extorted too, and they're asking for foreign aid for free lunches—what a disgrace!	negative
15	If the tax increases to 12%, the selling price must be raised to 110k without tax to make a profit. So the amount charged to consumers is 110k + 12% tax.	negative
16	The neighborhood shop isn't subject to the 12% tax? How could the regime give traditional retailers a break?	Negative
17	Do you agree? What's the reason? #tax #taxincrease #12%tax	negative
...	...	...
996	How can poverty decrease when taxes increase, layoffs are widespread, and unemployment is skyrocketing, sir? Many people are becoming migrant workers...	negative
997	When taxes increase by 12%, that's when poverty rates will rise, layoffs will become more widespread, prices of basic goods will skyrocket, purchasing power will decline, and unemployment will run rampant.	negative
998	The decline in purchasing power has ultimately affected the retail industry and F&B (food and beverage, etc.), which in turn has led to unemployment. Many of these unemployed people will eventually reduce tax revenue as well.	negative
999	Yes, that's correct, sir. Indonesian citizens are very happy with the tax increases. Despite the difficulty in finding new jobs and the high unemployment rate, the tax increases have not made life harder for us.	Negative
1000	Like an iceberg... mass unemployment will occur if the middle class, which accounts for 60% of Indonesia's economy, is affected by the tax increase.	Negative

The type of research method used is quantitative research. Quantitative research is research because the data process involves numbers[19] . The research data was collected from social media platform X (Twitter) using the links https://twitter.com . [20]Data analysis is a crucial procedure that transforms raw data into relevant and meaningful information by applying statistical or qualitative methods. The purpose of data analysis is to identify patterns, relationships, or trends in the data that can be utilized to address research problems or test theories. The success of the research and the validity of the findings depend on the selection of appropriate analysis procedures and the accurate interpretation of the analysis results.

2.6 Data Crawling (X)

In this study, the research data consists of comments from Indonesian citizens on X regarding the increase in taxes that led to rising unemployment rates from November 2024 to April 2025. The research dataset was obtained from a Python program written in Google Colab, as illustrated in Figure 3 below.

```
# Crawl Data

filename = 'dataa15.csv'
search_keyword = '"kenaikan" AND "pajak" since:2024-11-01 until:2025-04-28 lang:id'
limit = 1000

!npx -y tweet-harvest@2.6.1 -o "{filename}" -s "{search_keyword}" --tab "LATEST" -l {limit} --token {twitter_auth_token}
```

Figure 3. Data Crawling

The code above explains that data mining (crawling) was performed using the keywords 'kenaikan' and 'pajak' with comments written in Indonesian from November 1, 2024, to April 28, 2025. A total of 1,000 tweets will be crawled, and the crawling results will be exported to a file named 'dataa15.csv'.

3. RESULTS AND DISCUSSION

3.1 *SVM* Research Analysis

The implementation of Orange Data Mining[21] displays the design of the sentiment analysis widget interface using the *Support Vector Machine* algorithm integrated into the workflow, as shown in Figure 3 below:

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, Agustus 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8922
Hal 555-566
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

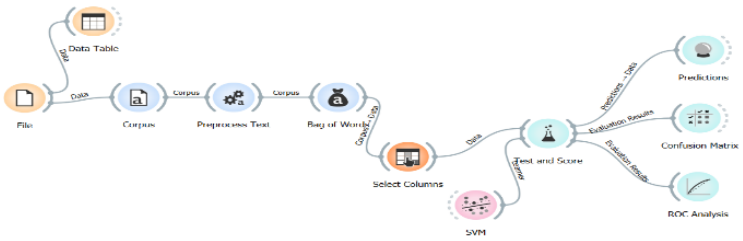

Figure 4. Support Vector Machine Workflow

Data collected (*crawled)* from social media platform X is inputted and analyzed individually based on its object. It is then linked to the necessary widgets for research purposes, resulting in a widget design as shown in the image above. Next, it enters the *text preprocessing* stage. Before text analysis, the text first undergoes preprocessing. This involves segmenting the text into smaller units (tokens), followed by transformation, tokenization, normalization, and filtering. Next, the process moves to the *Widget Select Columns* stage in Orange, which is used to select, filter, or modify attributes (columns) in the dataset before sending it to the next process or model. The process then proceeds to the *Test and Score* widget, as shown in the image below:

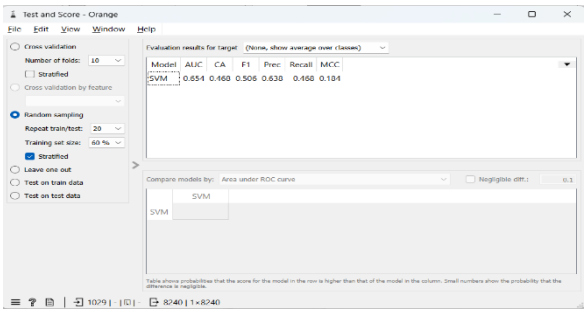

Figure 5. Test and Score

In the figure above, explain the selected evaluation method: Random Sampling. The model is evaluated by randomly dividing the data into training data (train) and testing data (test). Repeat train/test: 20 means that the train-test process is repeated 20 times to obtain more stable results. Training set size: 60% of the data is used to train the model, and the remaining 40% is used for testing. Results obtained:

Table 2. SVM Results

Metrics	Value	Description
AUC	0.654	Area Under the Curve – measures the model's ability to distinguish between classes. A value close to 1 is good; 0.5 = random. A value of 0.654 is still low.
CA	0.468	Total accuracy. Percentage of correct predictions (approximately 46.8%) is poor.
F1	0.506	Harmonic of precision and recall – suitable for imbalanced data
PREC	0.6	Of all positive predictions, how many are actually positive
RECALL	0.46	Of all the data that should be positive, how many were correctly recognized by the model?
MCC	0	Matthews Correlation Coefficient - indicates the strength of the correlation between the prediction and the actual value (values range from -1 to +1). A value of 0.184 indicates a very weak correlation.

The SVM model has low performance, as seen from: Accuracy below 50%, AUC only 0.654 → the model is only slightly better than random, low MCC → predictions are almost not disappointing compared to actual values This indicates that: The model is not yet able to distinguish well between sentiment classes (negative/neutral/positive).

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, Agustus 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8922
Hal 555-566
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

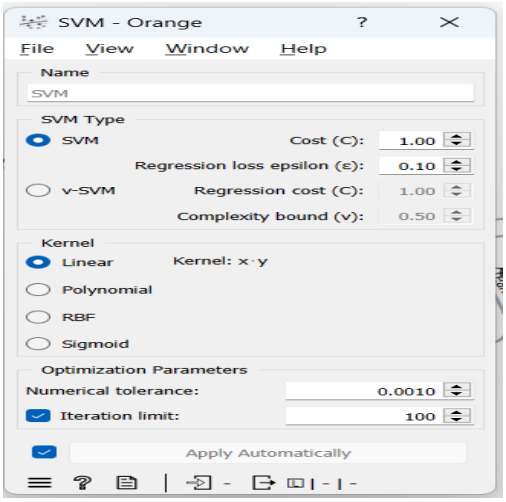

Figure 6. SVM display on Orange

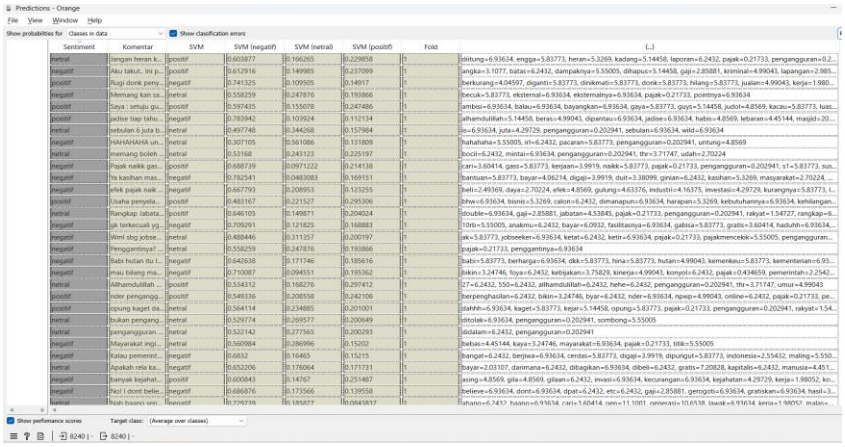

Figure 7. *Predictions* display

The first line of the image above explains the sentiment "neutral" (selected based on manual sentiment), the comment "Don't be surprised that sometimes people with such large balances don't report their taxes and are counted as unemployed" produces SVM "negative" with a value of SVM (negative) "0.603877" SVM (neutral) "0.166265" SVM (positive) "0.229858," and Fold "1" indicates that the data is in the first fold of the evaluation model (one of the parts of K-Fold Cross Validation).

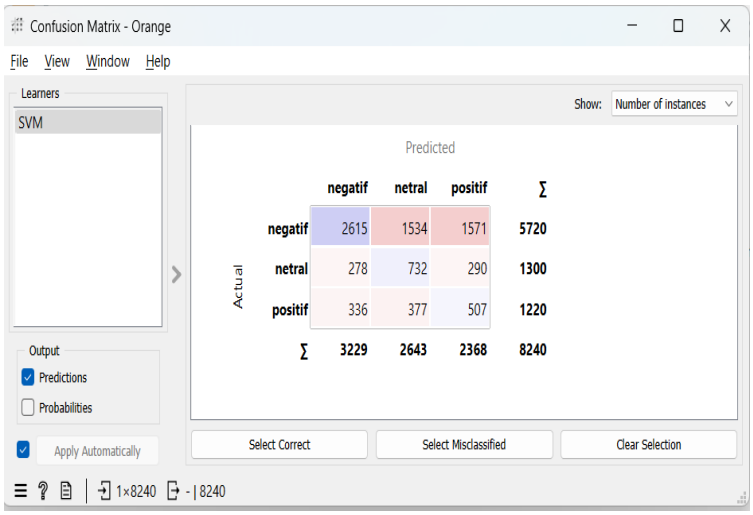

Figure 8. Confusion Matrix Display

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, Agustus 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8922
Hal 555-566
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Total Accuracy

$$Akurasi = \frac{\text{Number of correct predictions}}{\text{Number of data}} = \frac{2615 + 732 + 507}{8240} = \frac{3854}{8240} \approx 46{,}78\%$$

Precision, Recall, and F1-Score for each class (Negative, Neutral, Positive)

Table 3. Precision, Recall, and F1-Score

Class	Precision	Recall	F1-Score
Negative	0.8101	0.4573	0.5851
Neutral	0.2768	0.5631	0.3714
Positive	0.2141	0.4156	0.2823

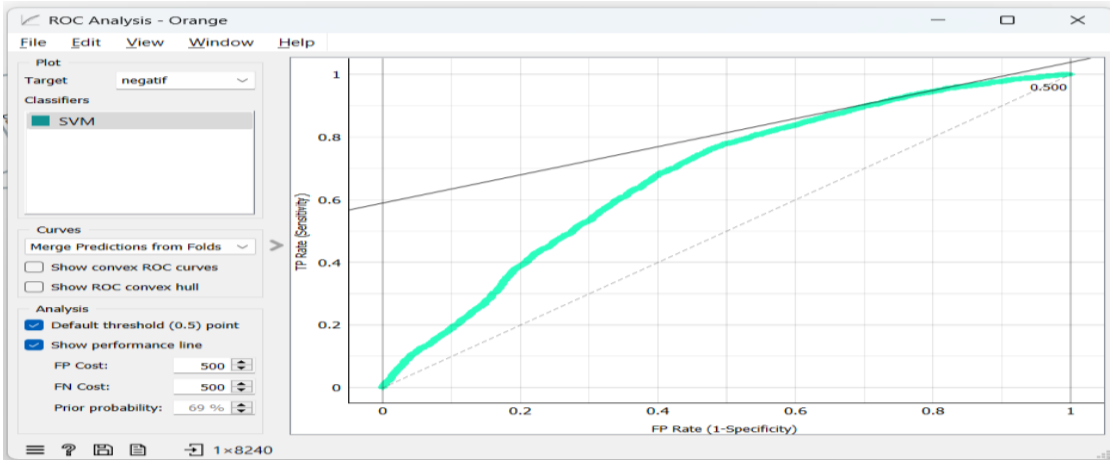

Figure 9. ROC Analysis Display for Target Class "negative"

The ROC (Receiver Operating Characteristic) curve for the SVM model with the target class "negative" shows the trade-off between sensitivity and specificity for various prediction thresholds. The more the curve curves upward to the left, the better the model performance. Target: The "negative" class is the positive class in this ROC context (binary classification for the ROC curve is done one-vs-rest). Curve Color: Light blue indicates the ROC curve for the SVM model. Default threshold: 0.5 is marked on the curve. Gray diagonal line: Baseline, representing a random model (without classification ability). AUC = 0.5. The ROC curve is nearly parallel to the diagonal line, meaning: The SVM model cannot distinguish well between the "negative" class and other classes. The AUC (Area Under Curve) value is likely to be close to 0.5, indicating nearly random or poor performance for the "negative" class. Default threshold (0.5): The threshold for determining whether an instance is predicted as "negative." FP Cost / FN Cost (500): The cost of errors for false positives and false negatives. Here, both are set equal. Prior probability: 69%: The proportion of the "negative" class in the dataset (approximately 69% of 8,240 total data points ≈ 5,720 negative data points).
AUC Comparison Table for SVM

Table 4. Comparison of AUC for Negative, Neutral, and Positive

Class	Prior Probability	AUC (Visual Estimate)	Interpretation
Negative	69	~0.58	Slightly better than random, the model recognizes this class fairly well due to its dominance
Neutral	16	~0.55	Poor, difficult to recognize, minority class
Positive	15	~0.53	Very weak, nearly random performance

3.2 Multinomial Naive Bayes Analysis

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, Agustus 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8922
Hal 555-566
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

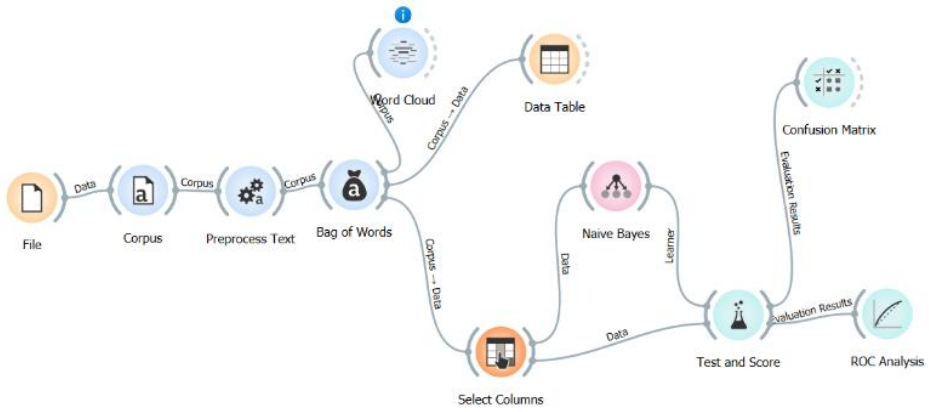

Figure 10. Multinomial Naive Bayes Workflow

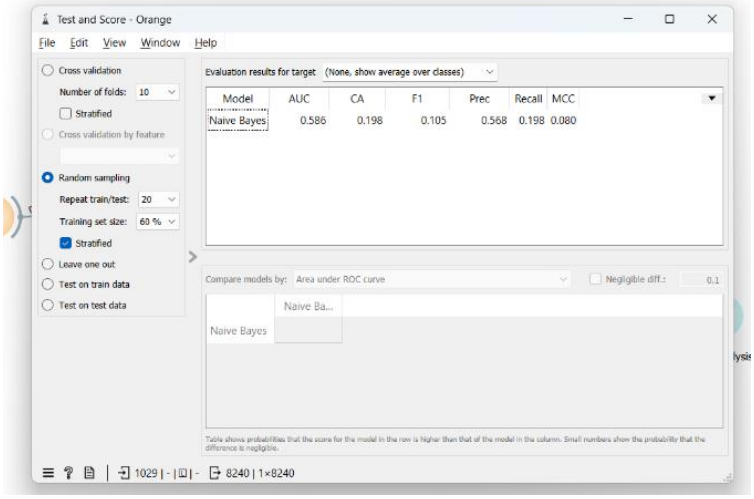

Figure 11. Test and score

Figure 11 is explained as follows:

Table 5. Test and Score

Metrics	Value	Description
AUC	0.586	Area Under Curve: Measures the model's ability to distinguish between classes. Scale 0–1. Values closer to 1 are better. Here, the value is low.
CA	0.19	Classification Accuracy: Only 19.8% of the data was predicted correctly. This is very low.
F1	0.105	F1-score: The harmonic mean of precision and recall. A value of 0.105 indicates that the model's performance is very poor.
Prec	0.56	Precision: Of all positive predictions, 56.8% are correct. This is average, but not too bad compared to other metrics.
Recall	0.198	Recall: Only 19.8% of positive cases were successfully found. Low.
MCC	0	Matthews Correlation Coefficient: An overall indicator of binary classification performance. A value close to 0 indicates random predictions.

Random sampling, model tested 20 times with random data division. The training set size is 60% of the data used for training, with the remainder used for testing. The class proportions (positive, negative, etc.) are kept balanced in the training and test data. Test on train/test data: Not selected, meaning the model is not tested only on the training data or only on the test data — instead, random sampling is performed repeatedly. The model has low accuracy (19.8%), very low F1-score and Recall, and nearly zero MCC, indicating very poor performance. Only Precision (56.8%) is still reasonable, but this is overshadowed by the extremely low Recall and F1 scores.

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, Agustus 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8922
Hal 555-566
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

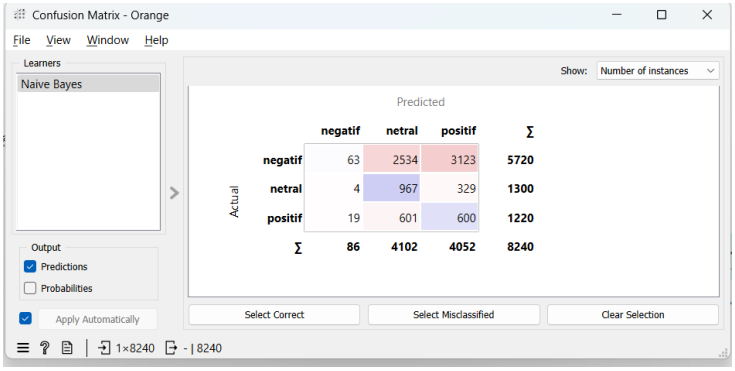

Figure 12. *Confusion matrix*

Table 6 Confusion matrix

	Prediction: Negative	Pred: Neutral	Prediction: Positive	Actual Number
Actual: Negative	63	2534	3123	5720
Current: Neutral	4	967	329	1300
Current: Positive	19	601	600	1,220
Predicted Total	86	4102	4052	8,240 (total)

The model accuracy is very low, as evidenced by the high number of misclassified data (e.g., only 63 out of 5,720 are correctly classified for the negative class). The Neutral class is the most frequently predicted (4102), indicating that the model tends to be biased toward the neutral class. This Confusion Matrix supports the poor metric values previously displayed in the *Test and Score* widget: low F1-score, low MCC, and CA only 0.198

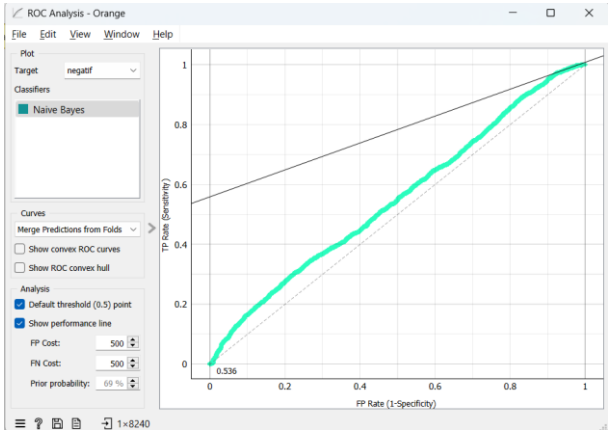

Figure 13 ROC Analysis for the 'negative' target class

Visualization of the ROC Curve (Receiver Operating Characteristic Curve) of the Naive Bayes model against the target class "negative" in Orange Data Mining. X-axis (FP Rate / 1 - Specificity): The further to the right, the more false positives. Y-axis (TP Rate / Sensitivity / Recall): The higher up, the more true positives. The ROC line (light blue/light green) is the result of the Naive Bayes model in distinguishing the negative class from others. This curve reflects the model's discrimination ability against the target class based on probability. The diagonal line (dashed) is the baseline or reference if the model makes random predictions. If the ROC curve approaches this line, it means the model is no better than random guessing. This indicates the model's performance at the default threshold (0.5). An AUC value of 0.536 means the model's performance is very weak, almost the same as random prediction. Ideal AUC = 1.0 (perfect), random = 0.5. Previous probability: 69%: This means that 69% of the data is negative class, indicating class performance (data imbalance). The Naive Bayes model is not effective in predicting negative classes. AUC 0.536 indicates accuracy close to random, which is not good enough to be used as the main classification model.

4. CONCLUSION

Based on the results of research and evaluation of sentiment classification models for three classes (negative, neutral, and positive) using the Support Vector Machine (SVM) and Multinomial Naive Bayes (MNB) algorithms, it was found that SVM consistently outperformed MNB. The evaluation was conducted using a random sampling approach with 20 repetitions, a training data proportion of 60%, and a test data proportion of 40%, along with stratification to maintain the

class distribution proportions. The SVM algorithm achieved an accuracy of 46.8%, F1-score of 0.506, precision of 0.638, recall of 0.468, AUC of 0.654, and MCC of 0.184. Meanwhile, the MNB algorithm only recorded an accuracy of 19.8%, F1-score 0.105, precision 0.568, recall 0.198, AUC of 0.586, and MCC of 0.080. The poor performance of MNB was caused by class imbalance in the dataset and the model's inability to recognize complex patterns between features, which are common limitations of the independence assumption in Naive Bayes. From these results, it can be concluded that the SVM algorithm is more reliable in sentiment classification because it provides more accurate, balanced, and relevant prediction results. It is concluded that SVM is more effective and accurate in classifying public sentiment regarding tax increases and their impact on poverty. Therefore, the SVM algorithm is recommended as a more appropriate approach for analyzing sentiment in the context of socio-economic issues such as this.

REFERENCES

[1] L. L. A. Mufida and M. S. Nasir, "Analisis Dinamis Tingkat Pengangguran di Indonesia," *J. Macroecon. Soc.*, vol. 1, no. 1, pp. 1–14, 2023.
[2] F. A. Tanjung, A. P. Windarto, and M. Fauzan, "Penerapan Metode K-Means Pada Pengelompokkan Pengangguran Di Indonesia," *Jurasik (Jurnal Ris. Sist. Inf. dan Tek. Inform.*, vol. 6, no. 1, p. 61, 2021.
[3] J. Han and M. Kamber, *Data Mining Concept and Technique*. San Francisco: Morgan Kauffman, 2006.
[4] M. Rafi, "Algoritma K-Means untuk Pengelompokan Topik Skripsi Mahasiswa," vol. 12, no. 2, pp. 121–129, 2020.
[5] A. C. Rahmat, "Identitas Proposal Penelitian," pp. 1–32, 2022.
[6] A. Ernawati, A. O. Sari, S. N. Sofyan, M. Iqbal, and R. F. W. Wijaya, "Implementasi Algoritma Naïve Bayes dalam Menganalisis Sentimen Review Pengguna Tokopedia pada Produk Kesehatan," *Bull. Inf. Technol.*, vol. 4, no. 4, pp. 533–543, 2023.
[7] W. B. Zulfikar, A. R. Atmadja, and S. F. Pratama, "Sentiment Analysis on Social Media Against Public Policy Using Multinomial Naive Bayes," *Sci. J. Informatics*, vol. 10, no. 1, pp. 25–34, 2023.
[8] Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, and Fitri Nurapriani, "Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN," *J. KomtekInfo*, vol. 10, pp. 1–7, 2023.
[9] M. Saputra, S. Nurhaliza Sofyan, A. Aulia, A. Ernawati, A. Oftasari, and R. Farta wijaya, "Implementation of E-Commerce System as SME Development Strategy in the Digital Era," *Bull. Inf. Technol.*, vol. 5, no. 3, pp. 195–202, 2024.
[10] Z. Sitorus, M. Saputra, S. N. Sofyan, U. Pembangunan, and P. Budi, "SENTIMENT ANALYSIS OF INDONESIAN COMMUNITY TOWARDS ELECTRIC," vol. 10, no. 1, pp. 108–113, 2024.
[11] Muit Sunjaya, Zulham Sitorus, Khairul, Muhammad Iqbal, and A.P.U Siahaan, "Analysis of machine learning approaches to determine online shopping ratings using naïve bayes and svm," *Int. J. Comput. Sci. Math. Eng.*, vol. 3, no. 1, pp. 7–16, 2024.
[12] M. Rasyid, Z. Sitorus, R. F. Wijaya, and M. Iqbal, "MACHINE LEARNING ANALYSIS IN IMPROVING THE EFFICIENCY OF THE STUDENT ADMISSION DECISION MAKING PROCESS NEW AT PANCA BUDI MEDAN DEVELOPMENT UNIVERSITY," vol. 3, no. 3, pp. 216–225, 2024.
[13] A. P. Nugraheni, S. N. Sunaningsih, and N. A. Khabibah, "Peran Konsultan Pajak Terhadap Kepatuhan Wajib Pajak," *Jati J. Akunt. Terap. Indones.*, vol. 4, no. 1, p. Editing, 2021.
[14] E. Suprapto, "Pengelompokkan Potensi Padi di Indonesia Menggunakan K-Means Cluster," *J. Ilm. Pop.*, vol. 5, no. 2, pp. 28–34, 2022.
[15] A. Pambudi and S. Suprapto, "Effect of Sentence Length in Sentiment Analysis Using Support Vector Machine and Convolutional Neural Network Method," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 15, no. 1, p. 21, 2021.
[16] R. Meiyanti, M. M. Munauwar, R. Fitria, and H. Al Kautsar, "Implementasi Algoritma K-Medoid pada Clustering Sayuran Unggulan di Kabupaten Aceh Utara," vol. 19, no. x, pp. 327–337, 2024.
[17] Y. S. T. Allo, V. Sofica, N. Hasan, and M. Septiani, "Penggunaan Metode Naïve Bayes Dalam Mengklasifikasi Pengangguran Pada Desa Bojong Kulur," *Bianglala Inform.*, vol. 10, no. 1, pp. 30–35, 2022.
[18] D. N. Ayu Ofta Sari, Muhammad Iqbal, "Analisis Faktor Demografi Dan Sosial Ekonomi Untuk Mendeteksi Dini Risiko Putus Kuliah Menggunakan Metode Support Vector Machine (Svm) Dan Decision Tree (Studi Kasus : STMIK Triguna Dharma)." JATILIMA, medan, p. 17, 2025.
[19] R. Millena and T. Jesi, "Jurnal Analisis Pendapatan Negara Indonesia Kota Bogor Provinsi Jawa Barat Dengan Metode Kuantitatif," *Jesya (Jurnal Ekon. Ekon. Syariah)*, vol. 4, no. 2, pp. 1004–1009, 2021.
[20] P. Candra Susanto, D. Ulfah Arini, L. Yuntina, J. Panatap Soehaditama, and N. Nuraeni, "Konsep Penelitian Kuantitatif: Populasi, Sampel, dan Analisis Data (Sebuah Tinjauan Pustaka)," *J. Ilmu Multidisplin*, vol. 3, no. 1, pp. 1–12, 2024.
[21] E. Mardiani *et al.*, "Membandingkan Algoritma Data Mining Dengan Tools Orange untuk Social Economy," *Digit. Transform. Technol.*, vol. 3, no. 2, pp. 686–693, 2023.