Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th October 2025. Vol. 5. No. 1; e-ISSN: 2808-4519

Public Sentiment Analysis on the Increase of Value Added Tax in Indonesia Through Tweet-Harvest

Hengky Triyo[1*],Aji Primajaya[2],Purwantoro[3]

[1,2,3]Informatics Engineering, Singaperbangsa University, Karawang
2110631170021@student.unsika.ac.id[1*], aji.primajaya@staff.unsika.ac.id[2], purwantoro.masbro@staff.unsika.ac.id[3]

Abstract.

The government's policy to increase the Value Added Tax (VAT) rate in 2025 has sparked various public reactions, particularly on the social media platform Twitter. This study aims to analyze public sentiment toward the policy using the Knowledge Discovery in Database (KDD) approach. Data were collected through Tweet-Harvest from January to May 2025 and processed through several stages, including text preprocessing, transformation into numerical representation using the Term Frequency–Inverse Document Frequency (TF-IDF) method, and feature selection with Information Gain. Sentiment classification was conducted using the Support Vector Machine (SVM) algorithm, while parameter tuning (hyperparameter tuning) was performed via Grid Search to optimize model performance. Model evaluation was carried out using accuracy, precision, recall, and F1-Score metrics. The analysis revealed that public opinions were categorized into three sentiment classes: positive, negative, and neutral, with negative sentiment being the most dominant. These findings provide insight into public perception of the VAT increase and can serve as a reference for the government in developing more effective and responsive policy communication strategies.

Keywords: *Sentiment Analysis, VAT, Twitter, SVM, Grid Search*

1 Introduction

Indonesia's Value Added Tax (VAT) increase policy in 2025 has elicited a variety of responses from the public. Since the VAT rate was raised from 11% to 12%, numerous public discussions have emerged on various social media platforms, particularly Twitter. This platform is often used by the public to voice opinions, both in support of and criticism of government policies. This situation demonstrates the importance of understanding public sentiment so that policy implementation can be more effective and gain public acceptance.

Sentiment analysis is a branch of text mining that focuses on extracting public opinion from text-based data. Previous research has shown that machine learning algorithms, particularly Support Vector Machines (SVMs), have a high ability to classify sentiment from social media. Several studies examining public responses to issues such as the COVID-19 vaccine, restrictions on public activities, and fuel price increases have demonstrated the effectiveness of SVMs in identifying positive, negative, and neutral sentiment. Based on these research findings, this study was conducted to examine public perception of the Value Added Tax (VAT) increase policy in Indonesia.

This research utilized the Knowledge Discovery in Database (KDD) method, which encompasses data selection, preprocessing, transformation, data mining, and evaluation. Data was collected using Tweet-Harvest with keywords related to VAT, then processed using text preprocessing and TF-IDF transformation techniques. Next, the SVM algorithm was applied to sentiment classification by optimizing parameters using Grid Search.

2 Literature Review

Sentiment Analysis in Public Opinion
Sentiment analysis has been widely used to understand public opinion regarding social issues and government policies. Through this analysis, data gathered from social media, especially Twitter, can provide insights into public perception of a particular policy or social phenomenon. Several previous studies have served as foundational references in this research, providing a basis for developing a more systematic and efficient approach.
Aisyah (2023)
Aisyah (2023) conducted sentiment analysis on Twitter data using the Support Vector Machine (SVM) algorithm with a lexicon-based approach. The results indicated that combining effective preprocessing methods with Term Frequency-Inverse Document Frequency (TF-IDF) weighting significantly improved classification accuracy. This study concluded that the combination of proper preprocessing and TF-IDF greatly contributed to enhancing the performance of sentiment classification models.
Putri and Idris (2024)
Putri and Idris (2024) studied public sentiment toward taxation policies using machine learning techniques. Their research emphasized the importance of optimal data preprocessing and hyperparameter optimization to achieve more accurate classification results, as well as to balance the class distribution. They found that improving the quality of preprocessing and performing hyperparameter tuning helped reduce class imbalance and improved the accuracy of sentiment analysis across positive, negative, and neutral categories.
Xu et al. (2022)
Xu et al. (2022), in their systematic review, explained that using social media as a data source for sentiment analysis offers significant opportunities for public policy research. However, they highlighted the challenges associated with data imbalance and the variation in language used by social media users. These factors require careful handling of data processing and sentiment classification, as tweets often contain informal language and diverse linguistic styles.

3 Research methods

This study applies the Knowledge Discovery in Database (KDD) approach, which consists of five main stages: data selection, preprocessing, transformation, data mining, and evaluation. The object of this research is public opinion regarding the planned Value-Added Tax (VAT) increase in Indonesia, collected from social media Twitter. Data is gathered using Tweet-Harvest, a web scraping tool that extracts tweets based on specific keywords. The data collection period is set from January to May 2025, with a focus on Indonesian-language tweets relevant to the research topic.

In the data selection phase, tweets related to the VAT increase topic are filtered using relevant keywords such as "kenaikan PPN" (VAT increase), "Pajak Pertambahan Nilai" (Value-Added Tax), and "PPN 2025". This selection process ensures that only tweets directly related to the VAT increase policy are included, providing a focused dataset for subsequent analysis.

Once the data is collected, the next stage is preprocessing, which is crucial for preparing the text data for analysis. In this phase, data cleaning is performed by removing irrelevant elements such as URLs, mentions (@user), hashtags, and emoticons, which may interfere with the analysis. Additionally, case folding is applied to convert all text to lowercase to ensure consistency, and **to**kenization is used to split the text into individual words or tokens. The preprocessing stage also involves stopword removal, eliminating common but non-informative words such as "dan" (and), "atau" (or), and "dari" (from). Stemming is applied to reduce words to their root form using the **S**astrawi library, which is specifically designed for Indonesian language processing. Lastly, normalization is performed to standardize terms and correct spelling errors.

After preprocessing, the next step is data transformation, where the cleaned text data is converted into a numerical format suitable for machine learning algorithms. This transformation is done using the Term Frequency-Inverse Document Frequency (TF-IDF) method, which evaluates the importance of each word in the corpus relative to the entire dataset. By applying TF-IDF, more weight is given to words that are rare across documents but frequent within a specific document, making these words more relevant for sentiment analysis. Additionally, feature selection is performed using the Information Gain algorithm to choose the most significant features that enhance sentiment classification. This step helps reduce dimensionality and improves the relevance and quality of features used in the model.

In the data mining phase, this study utilizes the Support Vector Machine (SVM) algorithm as the primary method for sentiment classification. SVM is chosen due to its proven effectiveness in analyzing natural language texts and its ability to separate data into distinct classes, in this case, positive, negative, and neutral sentiments. To maximize the model's performance, hyperparameter tuning is conducted using Grid Search, a technique that systematically tests different combinations of parameters such as kernel function selection (e.g., linear or radial basis function) and regularization parameter (C). This process optimizes the SVM model to ensure the best performance. Additionally, **cross-**validation is applied to ensure the model generalizes well and avoids overfitting the training data.

In the evaluation stage, the trained SVM model is tested using various metrics to assess its performance. The evaluation metrics include accuracy, which measures the proportion of correctly classified tweets out of the total number of tweets; precision, which calculates how accurately positive sentiment is classified; recal**l**, which measures how well the model detects positive sentiment; and the F1-Score, which provides a balance between precision and recall. A confusion matrix is also applied to analyze the model's ability to classify tweets into the three sentiment categories: positive, negative, and neutral. The confusion matrix will provide a detailed view of classification performance and help evaluate if the model is distinguishing between the classes effectively.

All of the research processes, including data collection, preprocessing, feature extraction, classification, and evaluation, will be implemented using the Python programming language on the Google Colaboratory platform. Google Colaboratory offers GPU support, enabling faster computation and model training. Python libraries such as Scikit-learn**,** Sastrawi, and Pandas will be used for data manipulation, machine learning, and text processing.

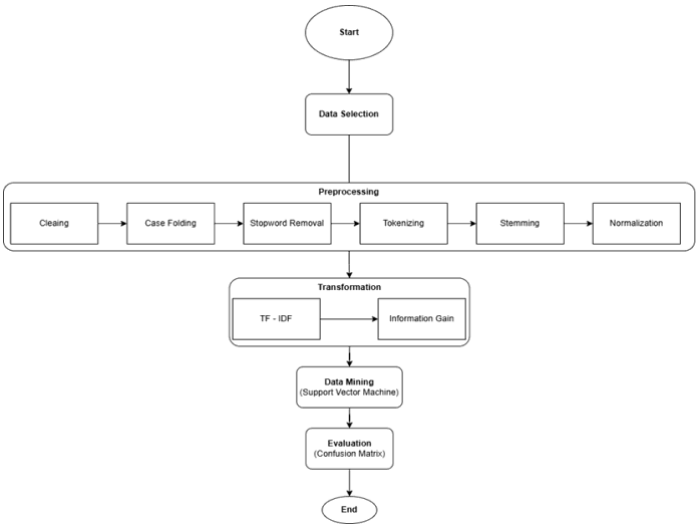

Figure 1: Flowchart of design flow.

4 Results and Discussion

The final evaluation of the model using the Support Vector Machine (SVM) algorithm, optimized through Grid Search, yielded an accuracy rate of 57.96% in classifying public sentiment regarding the Value-Added Tax (VAT) increase policy in Indonesia. A total of 2,161 tweets were analyzed, with the following sentiment distribution:

a. 374 tweets with positive sentiment
b. 1,261 tweets with neutral sentiment
c. 526 tweets with negative sentiment

The results indicate that neutral sentiment predominates in the public conversation surrounding the VAT increase issue, followed by negative and positive opinions in smaller proportions. The majority of public responses were neutral, suggesting that many individuals either did not have a strong opinion or did not express an emotional response towards the VAT increase policy. This neutrality might indicate that the public is either cautiously observing the policy changes or feels indifferent to the matter, rather than expressing strong support or opposition.

Neutral Sentiment Dominance

The dominant neutral sentiment suggests that many people did not feel strongly either in favor of or against the VAT increase. This could imply a lack of in-depth understanding or a desire for more detailed information about the policy. A neutral response may also reflect public

uncertainty or confusion about how the policy might impact them directly. In some cases, people may not have formed a firm opinion yet, pending further clarification or additional context from the government or relevant stakeholders.

Negative Sentiment

While neutral sentiment was dominant, negative sentiment was the second most prevalent. This indicates that a significant portion of the public harbors concerns or disapproval about the VAT increase. The negative sentiment could stem from concerns over the potential economic impact, such as increased living costs, reduced purchasing power, or skepticism about the effectiveness of the policy. People may also associate higher taxes with economic hardship, especially during times of economic uncertainty. Additionally, negative sentiment could reflect the frustrations of individuals who feel that the government's decision was made without sufficient consultation or communication.

Positive Sentiment

The positive sentiment was the least prevalent, which could indicate that, while some people may view the VAT increase as a necessary step for boosting government revenue, this perspective was less commonly expressed. Positive sentiment towards tax increases is often more difficult to generate, as tax hikes are typically seen unfavorably by the general public. However, some individuals may believe that the increase is justified, possibly due to its potential to improve public services or address economic imbalances.

Sentiment Distribution Results

The Sentiment Distribution after Preprocessing (as shown in the bar chart below) reveals that the neutral sentiment dominates the dataset, with over 1,200 tweets categorized as neutral, compared to about 500 negative and 350 positive tweets. This highlights that the public discussion around the VAT increase is largely cautious or non-committal.

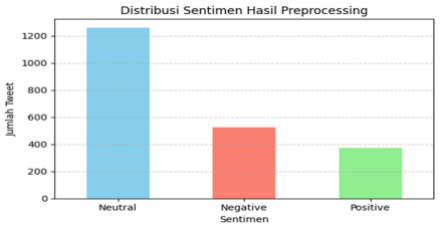

Figure 1: Sentiment Distribution after Preprocessing

Classification Report

In terms of performance, the model's precision, recall, and F1-Score were calculated and presented in the classification report shown below. The model performed well in identifying neutral sentiment, with a recall of 0.98 and an F1-Score of 0.73. However, the model struggled with classifying positive and negative sentiments, with precision and recall values of 0.00 for the positive sentiment class. This indicates that the model had difficulty distinguishing positive sentiment from the other classes, which may be due to the relatively low number of positive tweets and the challenges of accurately identifying positive opinions in social media data.

```
Test Accuracy: 0.5796766743648961

Classification Report:
              precision    recall  f1-score   support

          -1       0.17      0.01      0.02       103
           0       0.59      0.98      0.73       255
           1       0.00      0.00      0.00        75
```

Figure 2: Classification Report of the Sentiment Analysis Model

Implications for Policy Communication

These findings highlight the importance of effective government communication strategies. The prevalence of neutral and negative sentiment indicates that there is a need for the government to better communicate the reasons for the VAT increase, its expected benefits, and how it will impact different sectors of society. The government could consider engaging in more detailed public consultations, providing clear explanations about how the revenue will be used, and addressing concerns about the potential burden on lower-income groups.

The strong presence of neutral sentiment suggests that the public may be open to receiving more information before forming a clear opinion. Therefore, communication efforts should aim to educate and clarify the potential advantages of the VAT increase while addressing the public's concerns in a transparent and understandable way.

Limitations of the Study

Although the model achieved an accuracy of 57.96%, it is important to acknowledge some limitations. The accuracy rate is relatively modest, indicating that there is room for improvement in the model's ability to classify sentiments accurately. The imbalance in sentiment distribution (with a high proportion of neutral sentiment) may have impacted the model's performance. Additionally, factors such as the language used in the tweets, variations in phrasing, or the presence of sarcasm or irony may have posed challenges for the SVM algorithm, especially in distinguishing between positive and negative sentiments.

Suggestions for Future Research

Future research could explore alternative machine learning algorithms, such as Random Forest, Naïve Bayes, or Deep Learning models, to improve classification accuracy. Additionally, incorporating a larger and more balanced dataset, including tweets from diverse demographic groups, could lead to a more comprehensive understanding of public sentiment. Another area for improvement would be enhancing the preprocessing stage to handle informal language, slang, and abbreviations more effectively, as these are common in social media texts.

Furthermore, investigating the specific concerns raised in negative tweets and examining whether they align with broader public discourse could provide more actionable insights for policymakers. Analyzing sentiment over time, especially after key announcements or clarifications from the government, might also reveal shifts in public opinion, which could guide future policy adjustments.

5 Conclusion

Based on the results of the sentiment analysis on public opinion regarding the Value-Added Tax (VAT) increase policy in Indonesia, using Twitter data from January to May 2025, the following conclusions can be drawn:

1. Data Collection
   The data was successfully obtained using Tweet Harvest with relevant keywords, and the tweets were combined into a single dataset containing thousands of tweets related to the topic of the VAT increase. The data collection process ensured that the dataset was focused on the key topic being studied.
2. Preprocessing Results
   The data preprocessing stage, which included cleaning, tokenizing, stopword removal, stemming, and sentiment classification using an expanded lexicon-based approach, resulted in three sentiment categories: positive, negative, and neutral. The preprocessing steps effectively prepared the data for sentiment analysis and helped categorize the opinions expressed in the tweets.
3. Sentiment Distribution
   The final results show that the majority of public sentiment was categorized as neutral, followed by negative sentiment, and a smaller proportion of positive sentiment. This indicates that, while many individuals expressed their views, the majority of responses were either informational or did not strongly support or oppose the policy. The neutral sentiment may reflect a lack of strong opinion or emotional response towards the VAT increase, and possibly a wait-and-see approach from the public.
4. Model Performance
   The SVM algorithm, optimized using Grid Search, achieved an optimal parameter set with a test accuracy of approximately 57.96%. Although this accuracy is not considered high, the model was able to identify sentiment patterns quite well for the neutral class. However, it struggled with distinguishing between positive and negative sentiments, indicating that further improvement is needed in distinguishing these sentiments more accurately.
5. Implications of Findings
   The findings highlight that public opinion on social media regarding the VAT increase is diverse, and sentiment analysis serves as a valuable approach for quickly understanding public perceptions based on data. Despite the challenges in accurately classifying positive and negative sentiments, sentiment analysis provides useful insights into public sentiment trends, which can inform policy communication and strategy. The results underscore the importance of clear and transparent government communication to address public concerns and to improve understanding of policy changes among the population.

Reference

[1] Anggreini, N. M., "Pemanfaatan Media Sosial Twitter di Kalangan Pelajar SMK Negeri 5 Samarinda," *eJournal Sosiatri-Sosiologi, 4*(2), 2016. Available: https://ejournal.ps.fisip-unmul.ac.id/site/wp-content/uploads/2016/06/02_format_artikel_ejournal_mulai_hlm_genap-1%20(06-16-16-07-26-19).pdf

[2] Fitriyyah, S. N. J., Safriadi, N., & Pratama, E. E., "Analisis Sentimen Calon Presiden Indonesia 2019 dari Media Sosial Twitter Menggunakan Metode Naive Bayes," *Jurnal Edukasi dan Penelitian Informatika (JEPIN), 5*(3), pp. 279, 2019. Available: https://jurnal.untan.ac.id/index.php/jepin/article/view/34368/75676584400

[3] Fahmi, D. Y., Hartoyo, & Zulbainarni, N., "Mining Social Media (Twitter) Data for Corporate Image Analysis: A Case Study in the Indonesian Mining Industry," *Journal of Physics: Conference Series, 1811*(1), 012107, 2021. Available: https://www.researchgate.net/publication/350333146_Mining_Social_Media_Twitter_Data_for_Corporate_Image_Analysis_A_Case_Study_in_the_Indonesian_Mining_Industry

[4] Lim, S. A., "Implementasi Pajak Pertambahan Nilai di Indonesia: Suatu Studi Perbandingan di Negara-Negara Asean-9," *BIP's Jurnal Bisnis Perspektif, 12*(1), pp. 27-46, 2020. ISSN 1979-4932. Available: https://repository.ubaya.ac.id/43818/

[5] Yani R, Simandalahi E, Nasution A., "Pengaruh PPN (Pajak Pertambahan Nilai) terhadap Pendapatan Nasional Eksis," *Jurnal Ilmiah Ekonomi dan Bisnis, 15*(1), 30, 2024. Available: https://eksis.unbari.ac.id/index.php/EKSIS/article/view/424/208

[6] Liyana, N. F., "Menelaah Rencana Kenaikan Tarif PPN Berdasarkan Bukti Empiris serta Dampaknya Secara Makro Ekonomi," *Jurnal Pajak Indonesia (Indonesian Tax Review)*, 2021. Available: https://www.researchgate.net/publication/356996285_Menelaah_Rencana_Kenaikan_Tarif_PPN_Berdasarkan_Bukti_Empiris_Serta_Dampaknya_Secara_Makro_Ekonomi

[7] Pasek, P., Mahawardana, O., Arya, G., & Pratama, E., "Analisis Sentimen Berdasarkan Opini dari Media Sosial Twitter terhadap 'Figure Pemimpin' Menggunakan Python," *JITTER: Jurnal Ilmiah Teknologi dan Komputer, 3*(1), 2022. Available: https://www.neliti.com/publications/432979/analisis-sentimen-berdasarkan-opini-dari-media-sosial-twitter-terhadap-figure-pe

[8] Raditia Vindua, Achmad Udin Zailani, "Analisis Sentimen Pemilu Indonesia Tahun 2024 Dari Media Sosial Twitter Menggunakan Python," *JURIKOM JURNAL RISET KOMPUTER*, 2023. Available: https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom/article/view/5945/3483

[9] Graff, M., Moctezuma, D., Miranda-Jiménez, S., & Téllez, E. S., "A Python Library for Exploratory Data Analysis on Twitter Data based on Tokens and Aggregated Origin-Destination Information," *arXiv preprint arXiv:2009.01826*, 2020.

[10] Junianto, S., Harimurti, F., & Suharno, S., "Pengaruh Inflasi, Nilai Tukar Rupiah, Suku Bunga dan Self Assessment System terhadap Penerimaan Pajak Pertambahan Nilai di Kantor Wilayah Direktorat Jenderal Pajak Jawa Tengah II," *Jurnal Akuntansi dan Sistem Teknologi Informasi, 16*(1), 2023. Available: https://www.researchgate.net/publication/367590634_PENGARUH_INFLASI_NILAI_TUKAR_RUPIAH_SUKU_BUNGA_DAN_SELF_ASSESSMENT_SYSTEM_TERHADAP_PENERIMAAN_PAJAK_PERTAMBAHAN_NILAI_DI_KANTOR_WILAYAH_DIREKTORAT_JENDRAL_PAJAK_JAWA_TENGAH_II