Comparison Of Machine Learning Algorithms In Public Sentiment Analysis Of TAPERA Policy

Eklesia Sihombing[1*], Muhammad Halmi Dar[2], Fitri Aini Nasution[3]

[1,2,3] Faculty of Science and Technology, Universitas Labuhanbatu, Sumatera Utara Indonesia
[*]Corresponding Author:
Email: eklesia.hombing@gmail.com

Abstract.

The rapid development of information technology has changed the way people interact and express their opinions on public policies, including the People's Housing Savings (Tapera) policy in Indonesia. People now primarily express their views openly on social media platforms like Twitter, generating a substantial amount of text data for analysis to understand public sentiment. However, the main challenge in this sentiment analysis is determining the most effective machine learning algorithm for classifying public opinion with high accuracy. This study aims to compare the performance of three machine learning algorithms, namely Naïve Bayes, Support Vector Machine, and Random Forest, in analyzing public sentiment towards the Tapera policy. This study analyzes public comment data obtained from Twitter. We measure the accuracy of each algorithm to determine its optimal performance in sentiment classification. The research method consists of several stages, starting with data collection, text preprocessing to clean and prepare data, and then applying the three algorithms to analyze sentiment. The results showed that Naïve Bayes had the highest accuracy of 69.17%, followed by Support Vector Machine with an accuracy of 68.42%, and Random Forest with an accuracy of 66.17%. This shows that Naïve Bayes is the most effective algorithm to use in sentiment analysis of public comments related to the Tapera policy, especially in the context of complex text data from social media. The conclusion of this study is that Naïve Bayes is superior in classifying public sentiment towards the Tapera policy compared to Support Vector Machine and Random Forest. As a result, this study makes a significant contribution to selecting the most appropriate machine learning algorithm for public sentiment analysis towards public policy, which in turn can help the government understand and respond to public perceptions more effectively.

Keywords: *Machine Learning, Naïve Bayes, Random Forest, and Sentiment Analysis Support Vector Machine and Tapera.*

I. INTRODUCTION

Sentiment analysis in social media has become increasingly important in understanding public opinion and behavior [1], [2]. Social media data such as Facebook, Twitter, and Instagram can provide valuable insights for businesses and researchers in understanding consumer preferences, market trends, and competitor strategies [3]. It is important for businesses to use the right data analytics tools, take action on the insights gained, and pay attention to their privacy and data security policies [4]. Sentiment analysis can help identify positive, negative, or neutral attitudes in text such as reviews, comments, and social media posts [5], [6], [7]. Sentiment analysis can also offer recommendations for product enhancement, aid in the planning of a business startup, and identify sentiment in user comments [8]. Artificial intelligence techniques, such as natural language processing (NLP) technology, enable sentiment analysis in social media, including Twitter, by identifying and classifying data based on user sentiment towards a specific topic, product, or brand [9]. Sentiment analysis can provide valuable insights for companies to improve their marketing and customer service strategies [10]. Companies can use this information to inform their marketing strategies, enhance customer service, and monitor public perception over time [11]. Sentiment analysis has become an important tool in understanding people's feelings on a variety of topics, including public housing savings programs.

The policy on People's Housing Savings (Tapera) has raised pros and cons among the community. Tapera Fund management has not been fully effective in meeting Indonesia's housing financing needs. There are several challenges, such as low community participation, budget constraints, and lack of coordination between stakeholders [12]. The government must take into account the potential social impacts of the Tapera program, particularly its potential to exacerbate social and economic inequality, and the challenges faced by the informal sector and freelancers. The government needs to design a Tapera program that can reach all levels of society, including low-income groups [13]. To truly facilitate the community in having a decent home, Tapera's objectives and field implementation must synchronize [14]. Additionally, we need to enhance socialization, education, and coordination among stakeholders to effectively cater to the interests of independent workers in the Tapera program [15]. The Tapera program's implementation for private employees in Indonesia is critical. Low home ownership among private employees, limited access to housing

finance, and increasingly expensive house prices are urgent reasons for the government to implement this program [16].Many fields, including marketing, politics, and customer service, have widely used sentiment analysis. Sentiment analysis can help in understanding public opinion in a broad and deep way [17]. In social media, sentiment analysis is becoming increasingly relevant due to the large volume of data and its real-time nature. Research by [18] shows that Twitter can be a rich data source for sentiment analysis because users often share their opinions explicitly. Research by [19] shows that housing policies in Indonesia have a significant impact on people's welfare. Sentiment analysis of these policies can shed light on how people respond to them. According to [20], there is a dominant negative sentiment towards the Tapera policy and important issues such as financial burden, transparency, and policy effectiveness. The results of this study indicate a general dissatisfaction with the need to contribute, organized public protests, and how social media influences public discussion.

We have used a variety of machine learning algorithms for sentiment analysis. Naïve Bayes is one of the simplest and fastest algorithms. Researchers have successfully applied the Naïve Bayes algorithm to categorize positive and negative sentences, achieving accurate results on the matrix using the developed system [21]. Researchers [22], conducted a study that achieved the highest accuracy of the Naïve Bayes algorithm, 90.08%, using a 10% test data set. The Support Vector Machine (SVM) algorithm is known to have excellent performance in text classification [23]. The study conducted by [24] demonstrated that the SVM algorithm outperforms the Naive Bayes algorithm when it comes to airline reviews. Various studies have also proven the effectiveness of Random Forest, an ensemble method [25]. In user reviews of an e-commerce website, Random Forest gave the best accuracy of 97% compared to SVM [26]. This study aims to compare the performance of machine learning algorithms in sentiment analysis of public comments on the Tapera policy. We collected public comments from the Twitter social media platform. The algorithms to be compared are Naïve Bayes, Support Vector Machine, and Random Forest. We measure each algorithm's performance based on its accuracy. The best algorithm is the one that gets the highest accuracy score. We expect this study to offer recommendations for the most effective algorithms for sentiment analysis on this topic. This study is important because it can help improve understanding of how the public views the Tapera policy. Additionally, other researchers can use the study's results as a reference to develop further sentiment analysis methods. Therefore, we anticipate this study to make significant contributions to the fields of natural language processing and public policy analysis.

II. METHODS

This study compares the Naïve Bayes, SVM, and Random Forest algorithms in conducting sentiment analysis of public comments on Twitter regarding government policies related to People's Housing Savings (Tapera). In this study, the Python programming language is utilized on Google Collaborate online, from the process of reading the dataset to the visualization of the results. Researchers aim to simplify the algorithm implementation process by combining Google Colab with Python. Given the free nature of this application, installing the Python library is a straightforward process. Additionally, the Microsoft Excel application facilitates basic data visualization. Figure 1 shows the research steps taken.

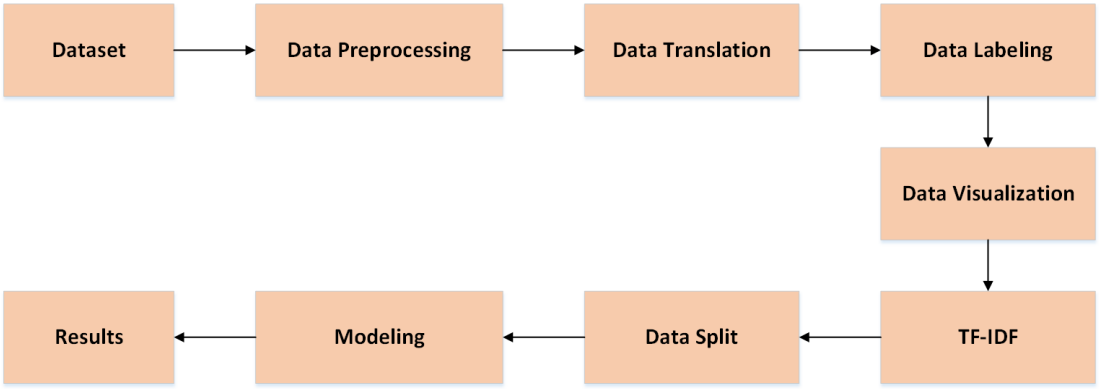

Fig 1. Research Stage

This study used secondary data from Kaggle [27]. This data consists of comments made by Indonesian-speaking individuals on Twitter about the Tapera policy. The preprocessing steps include removing punctuation, case folding, tokenizing, stopword removal, and stemming. Remove punctuation will rid the text of extraneous elements like punctuation, hashtags, emoticons, and so on. Case folding will convert all text to lowercase. Tokenizing is the process of breaking text into small units, called tokens. In sentiment analysis, stopword removal eliminates common words that lack significant meaning. Stemming is the process of changing words to their basic form. The collected tweets remain in Indonesian, necessitating a translation into English. Sentiment data needs to be translated into English because the sentiment model used to calculate the compound score in sentiment labeling is VADER (Valence Aware Dictionary and sEntiment Reasoner). The design and optimization of VADER for English encompasses the assessment of words, phrases, idioms, and cultural contexts unique to the language.

By translating the data into English, we can ensure that the model works optimally and provides accurate results. Lexicon-Based automates data labeling by calculating sentiment scores. Next, we visualize the labeling results in each sentiment class using diagrams and Wordcloud words to identify important and frequently discussed information. Term Frequency-Inverse Document Frequency (TF-IDF). TF will display the number of times a term appears in a document. IDF displays the quantity of documents that include the term. We use this method to ascertain the degree of significance a word holds within the completed document. Moreover, we separate the dataset into two parts: a training dataset and a testing dataset, with 90% of the data belonging to the former and 10% to the latter. At the modeling stage, three algorithms are implemented, namely, Naïve Bayes, SVM, and Random Forest. The test results will compare the accuracy scores of the three algorithms. Next, we will display the frequency distribution of the 10 most frequently appearing words.

III. RESULT AND DISCUSSION

This study uses a collection of tweets containing public opinions about the Tapera policy as its dataset. The analysis of this dataset aims to understand public sentiment and compare the effectiveness of various machine learning algorithms (Naïve Bayes, SVM, and Random Forest) in categorizing these sentiments. This study will provide an overview of how the public responds to the Tapera policy, which can be important input for policymakers.

Table 1. A part of Dataset

created_at	tweet (before preprocessing)	tweet (after preprocessing)
Tue Jun 18 00:16:25 +0000 2024	BP Tapera Tegaskan 124.960 Peserta Telah Ditindaklanjuti Dan Selesai. Mbappe Wisnu Bawah IPK 4 Daging Pagii https://t.co/zsvRpy9WeJ	bp tapera tegas serta ditindaklanjuti selesai mbappe wisnu ipk daging pagi
Mon Jun 17 23:59:09 +0000 2024	Semangat pagi warga tapera. Semangat beraktivitas hari ini .. https://t.co/MdHgr1JnXu	semangat pagi warga tapera semangat aktivitas
Mon Jun 17 23:47:17 +0000 2024	benci banget masalah tapera masalah pajak gaji aja udh emosi ini ditambah tapera 3% dr gaji. pengen nyundut pake roko deh. ga ada yg bkin emosi kecuali duit duit	benci banget tapera pajak gaji udh emosi tambah tapera gaji ken sundut pake roko bkin emosi kecuali duit duit

Table 1 is an excerpt from the Twitter user comment dataset consisting of 1189 records. The 'created_at' column indicates the date and time the user posted the tweet on Twitter. Meanwhile, the 'tweet' column encompasses the user's complete opinion or comments about the Tapera policy. The posting time range in this dataset is from 14 to 18 June 2024, indicating that the data was taken during a certain period that may be relevant to new announcements or policies related to Tapera. The significance of this date and time lies in its potential to reveal a sudden surge in either positive or negative sentiments, potentially linked to news or policy modifications. The tweets in this dataset show various public reactions to the Tapera policy. Table 1 also shows a comparison between the tweet text before and after preprocessing. Before the

preprocessing process, the "tweet" column contains the original tweet text. The text still contains elements such as punctuation, capital letters, links (URLs), and words that are irrelevant or uninformative for sentiment analysis.

The "tweet (after preprocessing)" column displays the tweet text after going through various preprocessing steps, such as punctuation removal, case folding, tokenizing, stopword removal, and stemming. The remove punctuation process eliminates punctuation like periods, commas, and exclamation marks. The machine learning algorithm simplifies the text to facilitate its processing. For instance, the first tweet eliminates all punctuation, leading to a more straightforward sentence. To avoid case sensitivity issues, the case folding process converts all characters to lowercase. For instance, the algorithm transforms the word "BP Tapera" into "bp Tapera." This is crucial because, without case folding, the algorithm may treat "Tapera" and "tapera" as two distinct words. The tokenizing process breaks down the tweet into individual tokens. This crucial step divides the text into discrete word units for independent analysis. The stopword removal process removes words considered unimportant for sentiment analysis, such as "and," "ini," and "untuk," from the text. For instance, sentiment analysis removes the words "ini" and "ada" from the third tweet because they don't provide important information. The stemming process transforms words into their fundamental forms. In the third tweet, the word "ditambah" becomes "tambah" and "nyundut" becomes "sundut." Stemming helps reduce variations of different words that have the same meaning, making analysis easier.

Table 2. Translation Results

Bahasa Indonesia	English
bp tapera tegas serta ditindaklanjuti selesai mbappe wisnu ipk daging pagi	bp tapera firm and followed up finished mbappe wisnu ipk morning meat
semangat pagi warga tapera semangat aktivitas	The morning spirit of Tapera residents is enthusiastic about activity
benci banget tapera pajak gaji udh emosi tambah tapera gaji ken sundut pake roko bkin emosi kecuali duit duit	I really hate the way the salary tax is already emotional

Table 2 shows the translation results from Indonesian text to English. This process is efficient because it only translates unique elements, thus reducing the number of calls to the translator API. This is important in the case of big data. In the context of sentiment analysis research, this translation plays a crucial role in enabling machine learning analysis on English text, a language that many machine learning models better support. The use of the googletrans library makes it easy to integrate the translation process directly into the data science pipeline without the need for separate pre-processing.

	Tweet	Compound_Score	Sentiment
0	bp tapera firm and followed up finished mbappe...	0.0000	Neutral
1	Bomb bayer heat	-0.4939	Negative
2	kt projo eitss finish blue	0.0000	Neutral
3	tapera order canang land bank faisal basri	0.0000	Neutral
4	Mum is complaining about the tapera issue, may...	-0.2023	Negative
5	Tapera leave is an example of command behavior...	0.3182	Positive
6	The morning spirit of Tapera residents is enth...	0.5994	Positive
7	It's really good for Indonesia, the salary is ...	0.4927	Positive
8	I really hate the way the salary tax is alread...	-0.4201	Negative
9	Minister loves the residents of Tapera Basuki ...	0.5719	Positive

Fig 2. Tweet Labeling Results based on Compound Score

Figure 2 shows the results of sentiment labeling using the VADER model on the dataset. Compound Score is a numeric value that indicates a comprehensive sentiment score for each text, ranging from -1 (very negative) to +1 (very positive). The sentiment label generated is based on the Compound Score value (positive, negative, or neutral). We designed VADER, a lexicon-based model, for sentiment analysis in

social media. VADER is able to handle informal language and the use of emoticons. It calculates a compound score based on the intensity of emotion in a text. VADER works by matching words in the text with an existing sentiment dictionary. The dictionary assigns a predetermined sentiment score to each word, and then adds these scores to create a compound score. Neutral sentiment is a compound score with a value of 0 (as in rows 0, 2, and 3) indicating that the text does not have a strong emotional tendency towards positive or negative. This can happen when the text contains factual information without a clear opinion. Negative sentiment is a negative compound score (e.g., -0.4939 in row 1, -0.2023 in row 4, or -0.4201 in row 8) that indicates the presence of negative sentiment in the text. For example, the texts "Bomb Bayer Heat" and "I really hate the way the salary tax is already emotional" explicitly contain negative expressions recognized by VADER. Positive sentiment is indicated by a positive compound score, such as 0.3182 in row 5, 0.5994 in row 6, and 0.5719 in row 9, which indicates the presence of positive sentiment in the text. Texts such as "The morning spirit of Tapera residents is enthusiastic about activity" and "Minister loves the residents of Tapera Basuki" are considered positive by the model.

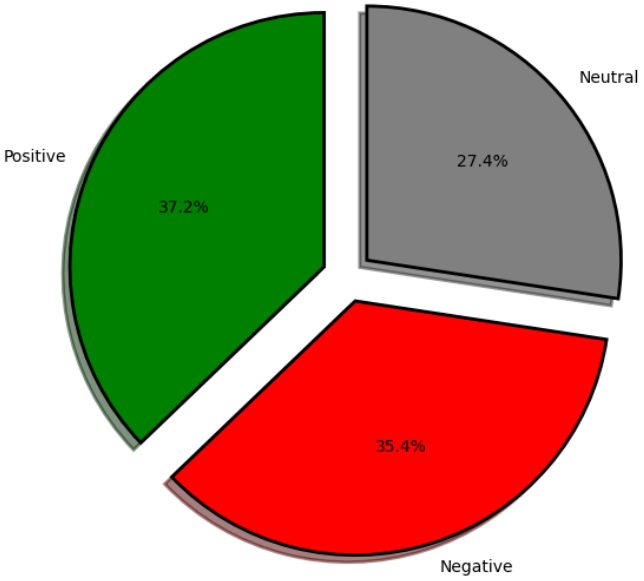

Fig 3. Distribution of Sentiment

Figure 3 shows the distribution of sentiment derived from 1189 tweets containing Twitter user reviews about the Tapera policy. Green represents the largest part of the pie chart, indicating that 37.2% (442 tweets) of the total tweets in the dataset have positive sentiment. This positive sentiment reflects an attitude or view that supports or feels optimistic about the Tapera policy. Red highlights the second largest portion, which accounts for 35.4% (421 tweets) of the total tweet count. This negative sentiment indicates that nearly one-third of the users who voiced their opinions about this policy expressed feelings of disagreement or dissatisfaction. Gray highlights the final segment, which accounts for 27.4% (326 tweets) of all tweets classified as neutral. This neutral sentiment means that the tweets do not show strong emotions, either positive or negative. Usually, these are more informative or descriptive tweets that do not include personal opinions. The difference between positive (37.2%) and negative (35.4%) sentiments is relatively small, suggesting a nearly equal divide in public opinion between those who support and those who reject the Tapera policy. However, there was a slight increase in positive sentiment, which may indicate that despite the criticism, there is still a fair amount of support for the policy. Neutral sentiment, which accounted for more than a quarter of the total tweets (27.4%), suggests that many people chose to express their opinions neutrally, or perhaps simply convey information related to the policy without expressing strong emotions. Overall, this distribution of sentiment suggests that the Tapera policy has elicited mixed reactions among the public. With almost half showing positive sentiment and the other half between negative and neutral sentiment, policymakers must further examine public opinion to identify areas where the policy could be improved or better communicated.

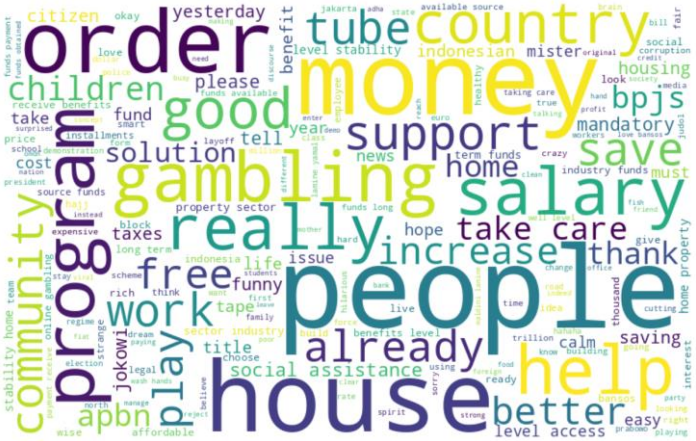

Fig 4. Positive Sentiment

Figure 4 shows a WordCloud showing the most frequently occurring words in tweets categorized as positive sentiment. The dominant words: "order," "program," "people," "money," "salary," "gambling," "country," "help," and "house" are some of the largest and most prominent words in this WordCloud. The large size of these words indicates that they appear most frequently in tweets with positive sentiment. The WordCloud reveals a positive perception of the Tapera policy, particularly when considering its financial and social implications. Many people feel that this policy can improve their welfare, both through financial support (such as salary and savings) and through social assistance and community support. Tapera's recognition as a relevant solution to the housing problem, a basic community need, is evident from the large number of words related to "house" and "home". Numerous terms such as "help," "support," and "community" suggest that people perceive this program not just as an economic policy, but also as a significant social endeavor.

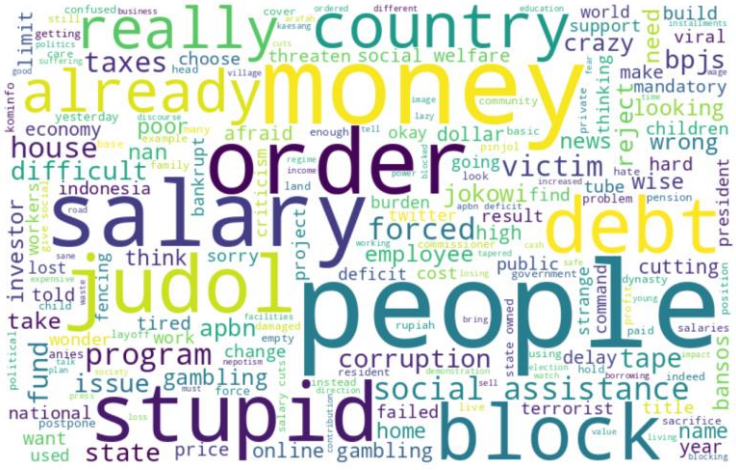

Fig 5. Negative Sentiment

Figure 5 is a WordCloud representation of the most frequently occurring words in tweets categorized as negative sentiment. Dominant words: "order," "people," "salary," "debt," "money," "country," "block," "stupid," "forced," and "judol" are some of the largest and most prominent words in this WordCloud. The large size of these words suggests that they frequently appear in tweets with negative sentiment. The WordCloud reveals that financial concerns, including debt burden, high costs, and negative impacts on salaries or income, primarily drive negative sentiment towards the Tapera policy. There are also social concerns related to the implementation of the policy, which is considered burdensome or even miserable for certain communities. Words like "forced," "corruption," and "criticism" suggest a robust opposition to the implementation of this policy, raising concerns about potential lack of transparency and unfairness in the process. The presence of words such as "block," "reject," and "Stupid" indicates that there are groups or individuals who strongly oppose this policy and may even be campaigning against it.

Fig 6. Neutral Sentiment

Figure 6 presents a visual representation of the WordCloud, highlighting the most frequently occurring words in the community's neutral sentiment towards the policy. The size of the words reflects their frequency of occurrence. "Tube" and "Order" are the most dominant words, their large appearance suggesting frequent discussions on these topics. This could indicate discussions about procedures or mechanisms related to the Tapera policy, such as how people order, use, or access related services. These three words (“house," “money," and “program”) also appear quite large, indicating that issues around housing, money, and the policy’s programs are major topics of community discussion. This may reflect a focus on how the Tapera policy affects people’s housing and finances. These words (“bank," “cost," and “salary”) indicate that there is significant attention to the financial aspects of the policy. People may be discussing how the policy affects costs, salaries, and the role of banks in its implementation. The terms "subsidized" and "assistance" suggest discussions about potential assistance or subsidies within the Tapera policy, as well as the community's implementation and reception of this assistance.

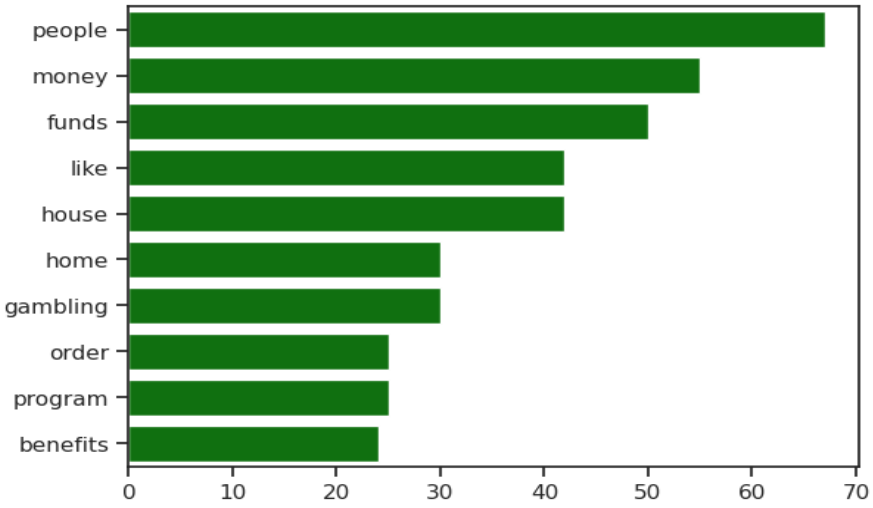

Fig 7. Frequency Distributions of 10 Most common words

Figure 7 shows the frequency distribution of the 10 most common words that appear in the analysis of public sentiment towards the Tapera policy. The word “people” appears 67 times; this word is the most frequent in the analysis, reflecting that public discussion or opinion in the context of Tapera is very focused on its impact on people or society. The word “money” appears 55 times, indicating that financial issues are the main topic in the discussion. The word "funds" is used 50 times, indicating that this policy's funding aspect is important. The words “like” and “house” (42 times each) most likely indicate a public preference or assessment of certain aspects of this policy. The word “home” (30 times) reinforces the focus on the housing aspect. The word “gambling” (30 times). The word "gambling" holds interest because it typically has no direct connection to housing policy. There may be discussions or concerns about the risks taken by the public or speculation related to the Tapera program. The words "order" and "program" appear 25 times each. The

term "order" might allude to the prescribed order or procedure in this policy. “Program” shows that the community also discussed the operational or design aspects of the Tapera policy itself. The word “benefits” (24 times), The relatively high frequency of "benefits" indicates a focus on what the community gets from this program. This word indicates that the community also discussed the benefits they expect or receive from the Tapera policy. The relatively high frequency of "benefits" indicates a focus on what the community gets from this program.

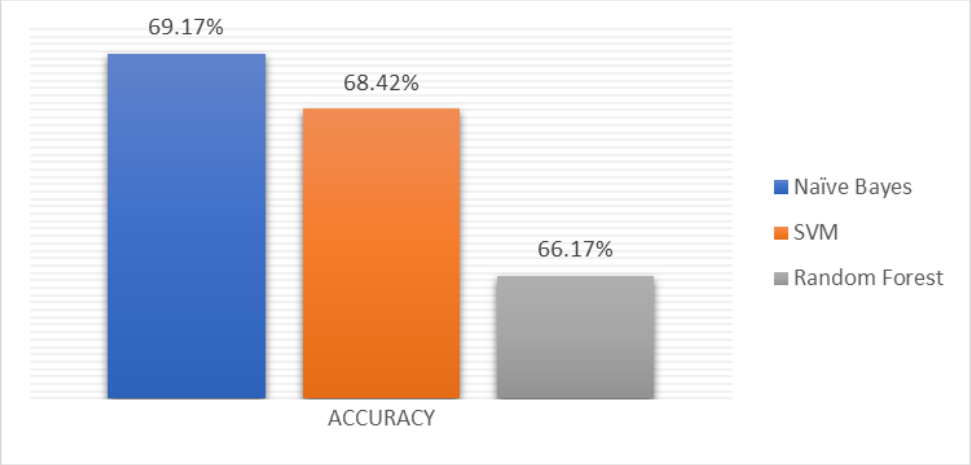

Fig 8. Comparison of Accuracy Results

Figure 8 shows a comparison of the accuracy results of three machine learning algorithms: Naive Bayes, Support Vector Machine (SVM), and Random Forest, in analyzing public sentiment towards the Tapera policy. The Naive Bayes algorithm has the highest accuracy results among the three algorithms, with 69.17% accuracy. This shows that Naive Bayes is quite effective in predicting public sentiment regarding the Tapera policy. Naive Bayes is known for its simplicity and computational efficiency, especially in cases of large data and independent features. In this context, Naive Bayes is able to classify sentiment well based on the distribution of words in the text. The SVM algorithm is in second place with an accuracy of 68.42%. Although slightly lower than Naive Bayes, SVM still shows excellent performance. SVM is usually very effective in finding a hyperplane that separates classes in the data with a maximum margin. Though SVM performs slightly worse than Naive Bayes, this may be due to data complexity or non-linear feature relationships. Random Forest has the lowest accuracy among the three algorithms, with a value of 66.17%. However, Random Forest still produces competitive results. Random Forest is usually good at handling data with non-linear relationships and reduces overfitting.

However, the lower accuracy in this study may indicate that this model is less effective at handling the specific text features used in this sentiment analysis.These results suggest that Naive Bayes may be better suited for use in text-based sentiment analysis, especially in a public policy context like Tapera, where word patterns may be fairly independent and less complex. SVM is still a strong alternative and can be considered when looking for a model with a clear classification margin, especially in cases with more linear data or highly correlated features. Although Random Forest performed less well, it can still be considered for other datasets or when model interpretability is important, as it can show which features are most influential in prediction. There may be benefits from further exploration of text data preprocessing, such as stemming, lemmatization, or the use of n-gram features, which could improve model accuracy. Combining results from multiple algorithms (e.g., using voting or stacking) may provide more accurate and stable results. Learning which features are most important to each model could provide further insight into how the public reacts to Tapera policies.

IV. CONCLUSION

This study has successfully compared the performance of three machine learning algorithms: Naïve Bayes, Support Vector Machine, and Random Forest in analyzing public sentiment on social media Twitter towards the People's Housing Savings (Tapera) policy. Based on the results of the study, Naïve Bayes is the algorithm that shows the best performance in this study, with the highest accuracy score of 69.17%. This

shows that Naïve Bayes is more effective in classifying public sentiment towards the Tapera policy compared to other algorithms tested. Support Vector Machine is in second place with an accuracy of 68.42%, only slightly below Naïve Bayes. Despite its lower accuracy, SVM is still a powerful and reliable algorithm for sentiment analysis tasks in this context. Random Forest shows the lowest performance with an accuracy of 66.17%. Nevertheless, this algorithm still produces significant results and can be used in situations where model interpretation or identification of important features is more important. From these results, it can be concluded that Naïve Bayes is the most appropriate algorithm for use in analyzing public sentiment towards the Tapera policy, at least in the context of the dataset and preprocessing used in this study.

Given the best performance shown by Naïve Bayes, the main suggestion is to use this algorithm for similar tasks in the future, especially for text sentiment analysis from social media. Naïve Bayes can also be used as a reliable baseline model to compare the performance of other algorithms in various sentiment analysis studies. Although Naïve Bayes gave the best results, further research can consider testing with other algorithms such as logistic regression, gradient boosting, or deep learning to see if any model can provide higher accuracy. We can test the use of ensemble techniques like stacking or bagging to see if a combination of several models can yield improved performance. According to the results of this study, the development of an application or system that can analyze sentiment in real-time from social media related to public policy can help the government or policymakers better understand public perception. You can also use this application to track changes in public sentiment over time following policy implementation. We hope that further research, taking into account these conclusions and suggestions, can further optimize public sentiment analysis and yield more accurate and useful insights for public policy decision-making.

REFERENCES

[1] Suwardono and G. Santoso, "Peran Media Massa dan Opini Publik dalam Mendukung atau Mengancam Kesehatan Demokrasi," ***J. Pendidik. Transform***., vol. 02, no. 03, pp. 239–249, 2023, doi: 10.9000/jpt.v2i3.1393.

[2] I. G. P. Megayasa, P. P. O. Mahawardana, and P. R. Nurbawa, "Analisis Sentimen berdasarkan Opini dari Media Sosial Twitter terhadap 'Figure Pemimpin' menggunakan Python,***" J. Manaj. dan Teknol. Inf.***, vol. 13, no. 1, 2023, doi: 10.5281/zenodo.7934336.

[3] A. Andirwan, V. Asmilita, M. Zhafran, A. Syaiful, and M. Beddu, "Strategi Pemasaran Digital: Inovasi untuk Maksimalkan Penjualan Produk Konsumen di Era Digital,***" JIMAT J. Ilm. Multidisiplin Amsir,*** vol. 2, no. 1, pp. 155–166, 2023, doi: 10.62861/jimat%20amsir.v2i1.405.

[4] U. Aulia and M. I. P. Nasution, "Memanfaatkan Data Media Sosial untuk Intelijen Kompetitif di Era Digital,***" J. Informatics Busisnes***, vol. 02, no. 01, pp. 78–83, 2024.

[5] O. Manullang, C. Prianto, and N. H. Harani, "Analisis Sentimen untuk Memprediksi Hasil Calon Pemilu Presiden menggunakan Lexicon Based dan Random Forest,***" J. Ilm. Inform.***, vol. 11, no. 02, pp. 159–169, 2023, doi: 10.33884/jif.v11i02.7987.

[6] N. Raisa, N. Riza, and W. I. Rahayu, "Analisis Sentimen menggunakan SVM dan KNN pada Review Drama Korea di MYDRAMALIST," ***JINTEKS (Jurnal Inform. Teknol. dan Sains)***, vol. 5, no. 4, 2023, doi: 10.51401/jinteks.v5i4.3114.

[7] D. Pramana, M. Afdal, Mustakim, and I. Permana, "Analisis Sentimen terhadap Pemindahan Ibu Kota Negara menggunakan Algoritma Naive Bayes Classifier dan K-Nearest Neightbors,***" J. Media Inform. Budidarma,*** vol. 7, no. 3, pp. 1306–1314, 2023, doi: 10.30865/mib.v7i3.6523.

[8] Z. Munawar *et al.*, *Big Data Analytics: Konsep, Implementasi, dan Aplikasi Terkini*, Pertama. Bandung: Kaizen Publisher, 2023.

[9] A. Kaharudin, A. A. Supriyadi, Muhlis, H. Baitika, and M. Derryanur, "Analisis Sentimen pada Media Sosial dengan Teknik Kecerdasan Buatan Naïve Bayes: Kajian Literatur Review," ***OKTAL J. Ilmu Komput. dan Sci.,*** vol. 2, no. 6, pp. 1642–1649, 2023, [Online]. Available: https://journal.mediapublikasi.id/index.php/oktal/article/view/2944%0Ahttps://journal.mediapublikasi.id/index.php/oktal/article/download/2944/1371

[10] B. H. Dzakiyyah, K. D. Putri, N. Y. Salsabila, T. A. Rafania, and I. F. A. Prawira, "Pemanfaatan Big Data untuk Meningkatkan Kepuasan Pelanggan Shopee," ***Innov. J. Soc. Sci. Res.,*** vol. 3, no. 5, pp. 10441–10455, 2023, doi: 10.31004/innovative.v3i5.5534.

[11] N. Ulyah, "Analisis Strategi Pemasaran untuk Meningkatkan Penjualan Pada PT. Bhirawa Steel," 2016. [Online]. Available: http://eprints.perbanas.ac.id/163/1/ARTIKEL ILMIAH.pdf

[12] B. A. Putri and R. Prijadi, "Public Fund Optimization for Housing Finance (Case Study: Tabungan Perumahan Rakyat, Indonesia)," in *Proceedings of the 5th International Conference on Economics, Business and Economic Education Science, ICE-BEES 2022*, 2023. doi: 10.4108/eai.9-8-2022.2338624.

[13] M. Ihsan, A. Rofiq, and Khusnudin, "Polemik Tabungan Perumahan Rakyat (Tapera): Sebuah kajian dengan pendekatan interdisipliner," ***Gulawentah J. Stud. Sos.,*** vol. 9, no. 1, pp. 72–86, 2024, doi: 10.25273/gulawentah.v9i1.20497.

[14] C. Ariningdyah, D. Lasonda, and F. R. D. Miarsa, "Analisis Yuridis Penerapan Tabungan Perumahan Rakyat (Tapera) dalam Perspektif Asas Keadilan," ***Innov. J. Soc. Sci. Res***., vol. 4, no. 3, pp. 18410–18424, 2024, doi: 10.31004/innovative.v4i3.12769.

[15] N. Haviazzahra, "Analisis Hukum Kepesertaan Pekerja Mandiri dalam Pelaksanaan Program Penyelenggaraan Tabungan Perumahan Rakyat," ***Aliansi J. Hukum, Pendidik. dan Sos. Hum.,*** vol. 1, no. 5, 2024, doi: 10.62383/aliansi.v1i5.386.

[16] M. Pasah, M. Yohana, and H. Winata, "Urgensi Penerapan Tapera bagi Pegawai Swasta di Indonesia**," *CAUSA J. Huk. dan Kewarganegaraan,*** vol. 5, no. 2, 2024, doi: 10.3783/causa.v2i9.2461.

[17] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, 2nd Editio. Cambridge University Press, 2020.

[18] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds., Valletta, Malta: European Language Resources Association (ELRA), May 2010. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf

[19] A. Suryahadi, R. Al Izzati, and D. Suryadarma, "The Impact of COVID-19 Outbreak on Poverty: An Estimation for Indonesia," 2020. [Online]. Available: http://smeru.or.id/en/content/impact-covid-19-outbreak-poverty-estimation-indonesia

[20] N. Rulandari, "Public Participation in Policy Making: Sentiment Analysis of TAPERA Policy on Twitter," ***Ilomata Int. J. Soc. Sci.,*** vol. 5, no. 3, 2024, doi: 10.61194/ijss.v5i3.1296.

[21] E. D. Harahap and R. Kurniawan, "Analisis Sentimen Komentar terhadap Kebijakan Pemerintah Mengenai Tabungan Perumahan Rakyat (TAPERA) pada Aplikasi X menggunakan Metode Naïve Bayes," ***J. Tek. Inform. Unika ST. Thomas***, vol. 09, no. 01, 2024.

[22] T. S. Rambe, M. N. S. Hasibuan, and M. H. Dar, "Sentiment Analysis of Beauty Product Applications using the Naïve Bayes Method," *SinkrOn*, vol. 8, no. 2, pp. 980–989, 2023, doi: 10.33395/sinkron.v8i2.12303.

[23] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," in *Machine Learning: ECML-98*, C. Nédellec and C. Rouveirol, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 137–142.

[24] A. M. Rahat, A. Kahir, and A. K. M. Masum, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset," in *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, 2019, pp. 266–270. doi: 10.1109/SMART46866.2019.9117512.

[25] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[26] P. Karthika, R. Murugeswari, and R. Manoranjithem, "Sentiment Analysis of Social Media Network Using Random Forest Algorithm," in *2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, 2019, pp. 1–5. doi: 10.1109/INCOS45849.2019.8951367.

[27] Kaggle, "Tapera Dataset," kaggle.com. Accessed: Jun. 25, 2024. [Online]. Available: https://www.kaggle.com/datasets/unshoytable/twitter-tapera-dateset.