Sentiment Analysis of Twitter Posts Using Naïve Bayes, Support Vector Machine, and Random Forest

Kevin
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
kevin094@binus.ac.id

Vincent Nicholas Hie
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
vincent.hie@binus.ac.id

Muhammad Fikri Hasani
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
muhammad.fikri003@binus.ac.id

Pandu Wicaksono
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
pandu.wicaksono005@binus.ac.id

***Abstract*—This paper looks into the effectiveness of three machine learning algorithms for sentiment analysis on Twitter posts: Random Forest, Support Vector Machine (SVM), and Naïve Bayes. Sentiment research may assist companies and content creators in customizing their approaches based on user sentiment and is essential for comprehending public opinion. This paper seeks to identify the best technique for classifying Twitter sentiment into positive, negative, and neutral categories by contrasting the accuracy and F1-score of different algorithms. This paper runs each algorithm on the preprocessed Twitter data comments, compares the results to a Kaggle dataset, and concludes with discussion. Significant variations in accuracy and runtime are shown by the findings: Although the Naive Bayes method is the quickest, it is also the least accurate, which means that huge and complex datasets are not the best fit for it. Both SVM and Random Forest are more accurate than Naïve Bayes with SVM slightly outperforms Random Forest in terms of accuracy. However, both models require greater processing power and longer training periods. Since SVM and Random Forest have higher accuracy and ability to handle noisy data, they would be more preferred in applications that depend on accuracy and strong resistance to noisy data though they are more computationally expensive. The present research's analysis presents a thorough and methodical comparison of these algorithms which are being used in sentiment analysis studies.**

Keywords*—*Twitter, Support Vector Machine, Naïve Bayes, Random Forest, Sentiment Analysis, Machine Learning

I. Introduction

With over 335,7 million active users [1], Twitter has grown to be one of the most widely used social networking platforms worldwide. Twitter allows users to express their feelings and thoughts through users' broadcast short posts. Twitter's diverse and large user base makes it a great market for content creators and businesses. Understanding the sentiment in Twitter posts will reveal a few key things. It shows how people perceive videos or products and reveals market preferences. Also, understanding sentiment helps engage audiences better. Creators and businesses can adjust their content to match customers' preferences with sentiment analysis, boosting engagement and impact. By studying sentiment analysis, content creators can adapt their content and messaging to better suit the desires of users, therefore increasing engagement and having a greater impact.

Sentiment analysis studies how to label texts based on feelings shown. This paper uses sentiment analysis for Twitter posts. The Naïve Bayes classification method is simple yet effective for text classification tasks. It can handle big datasets efficiently without needing lots of computing power. This makes it suitable for analyzing the large dataset of Twitter posts. Meanwhile, the Support Vector Machine classification method can be used for Twitter dataset because SVM has robustness to noise, where Twitter comments can be noisy with misspellings, slang, and abbreviations [2]. Random forest is suitable to be used, when the model has to balance between interpretability, performance, and the ability to handle complex relationships within the data. Naive Bayes, Random Forest and Support Vector Machine can determine if the sentiment is positive, neutral, or negative.

This study aims to comprehend the overall sentiment expressed in Twitter posts. Simultaneously, it evaluates the usability and precision of Support Vector Machine Machine, Naïve Bayes classifier, and Random Forest for sentiment analysis tasks. Our objectives are firstly, to enhance the understanding of digital sentiment dynamics, and secondly, to develop a tool that can provide insights and in-depth comprehension about videos and products for businesses, content creators, and researchers utilizing the Twitter platform. Thirdly, this paper wants to compare the effectiveness of both classifiers in the Twitter dataset. By leveraging machine learning algorithms, this study seeks to contribute to the advancement of sentiment analysis methodologies.

II. Related Work

In [3], a comparison between Support Vector Machine and Naïve Bayes algorithms in sentiment analysis, especially on Twitter data, is highly favored in recent research efforts. One of the researchers is Styawati et al. in 2021, which compares the effectiveness of both methods: Naïve Bayes and SVM in classifying sentiment in social media content.

SVM is known as a strong classification method with a high level of accuracy. In research conducted by Styawati et al. in 2021, SVM has an accuracy rate of 88.52% in classifying Twitter data, which means it is effective for sentimental analysis tasks. In addition, it is mentioned that SVM has been praised for its function of being able to find optimal hyperplanes, which separate data classes in an ideal way, resulting in superior performance in classifying tasks.

Naïve Bayes has also emerged as a formidable competitor in sentiment analysis. Even though SVM is considered superior in certain conditions, Naïve Bayes has proven good classification capabilities, especially compared to other methods such as K-NN. Research conducted by Styawati et al. in 2021 shows that Naïve Bayes has an accuracy rate of up to 75.58% in classifying tweet data.

In light of the above previous research, our research attempts to contribute to the ongoing discourse regarding sentiment analysis methodology by conducting a comparative analysis of SVM and NB algorithms. We seek to evaluate the performance of these methods in the accurate classification of sentiments expressed on Twitter. In addressing sentiment analysis in the context of social media conversation, our research aims to provide a meaningful contribution by providing insightful analysis of the advantages and disadvantages of utilizing SVM, Random Forest, and Naïve Bayes.

In [4], there is a study that examines the use of machine learning techniques, including Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) classifiers, for social media sentiment analysis and classification. This classifier has been successful in categorizing tweets and other social media posts into domains such as education, business, crime, and health. For example, Naw Naw. (2018) used SVM and KNN to analyze Twitter data, and successfully categorized posts to better understand public sentiment towards various social issues.

In [5], Studies on sentiment analysis that have recently been published, such those by Munir Ahmad, Shabib Aftab, and Iftikhar Ali, have looked closely at how well Support Vector Machine (SVM) models work with pre-classified tweet datasets. Weka was used to measure performance using datasets focused on self-driving cars and Apple devices. The study's findings revealed that the usefulness of the SVM differs dramatically between datasets, as seen by average precision, recall, and F-measure scores of 55.8%, 59.9%, and 57.2% for the first dataset and 70.2%, 71.2%, and 69.9% for the second. This performance variability emphasizes the role of dataset factors in shaping SVM outputs.

Sentiment analysis is the study of people's attitudes, sentiments, and emotions about specific things. In [6], this study addresses a key issue in sentiment analysis: categorizing sentiment polarity. This research uses tweets about six airlines from Kaggle.com. A random forest algorithm was used for sentiment analysis to reach an accuracy level of 75%.

In [7], there is a study that contributes to the data distribution work by evaluating algorithms using different training and test data splits (10%:90%, 20%:80%, 30%:70%, and 35%:65%). Consistent with previous findings, the Random Forest algorithm demonstrated superior accuracy and Area Under the Curve (AUC) metrics in most scenarios, specifically excelling at a 10%:90% split with an accuracy of 93.08% and an AUC of 0.962, and at 20%: 80% split with 93.45% accuracy and AUC 0.966. In contrast, for a 30%:70% split, Decision Tree shows comparable performance, while the Ensemble method is most effective at a 35%:65% split. Furthermore, this research highlights the important role that data preprocessing has in model performance, supporting the notion that meticulous preparation of the data set is essential to raising AUC values and prediction accuracy. In particular, these insights help to forecast user sentiment on social media sites like Twitter. These findings complement and broaden understanding of how different machine learning algorithms perform in sentiment analysis tasks.

III. Methodology

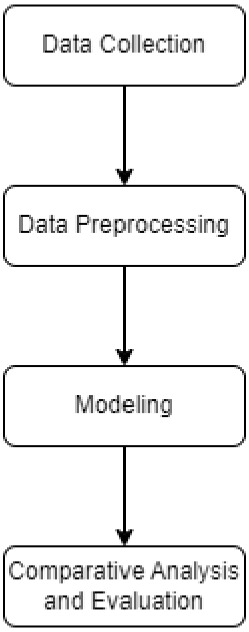

Figure 1 : Methodology Diagram

A. Data Collection

This paper uses quantitative research for the accuracy of sentiment classification. This paper uses kaggle for data collecting. The data sampled some twitter posts and labeled it with sentiment which consist positive, negative, and neutral. There's two dataset in kaggle such as "training" and "validation". Dataset training used to train machine learning and validation to check the accuracy of the algorithm. The data consist of four columns namely tweetID, entity, sentiment, tweetContent.

B. Data preprocessing

There is still redundancy in the dataset. Duplicate and null value in column has to be removed. And columns that are being used for modeling sentiment and tweetContent so entity and tweetID have to be removed. The text in the dataset can be improved by performing tokenization, removing stop words, and stemming.

C. Modeling

In this paper, the algorithms that will be used are Naïve Bayes, Support Vector Machine, and Random Forest. All algorithms excel at classification of two variables.

1. **Naïve Bayes**

   The Naïve Bayes Classifier is a classification method based on Bayes' Theorem, which calculates the probability of a data point belonging to a certain class given its features [8]. Probability theory is used for classification tasks. Naïve Bayes sentiment analysis uses the probabilities found in a text to determine the probability that a text belongs to a particular class of emotions (positive, negative, or neutral).

   Naïve Bayes is sometimes called a probabilistic classifier because it is based on the Bayes Theorem.. The type of naive bayes that is commonly used for sentiment analysis is Multinomial Naïve Bayes (MNB) because it is efficient at handling big data and performs well in high-dimensional spaces.

2. **Support Vector Machine (SVM)**

   Support Vector Machine (SVM) has been an important machine learning algorithm in recent times due to its unique merits for handling small sample sizes, nonlinear patterns, and high-dimensional data. Based on the theory of Statistical Learning Theory (SLT), SVM boasts complete theoretical clarity, global optimization, strong adaptability, and excellent generality. The algorithm works by finding the optimal hyperplane that best separates different classes of data points in the feature space. [9]

   There's two methods in Support Vector Machine such as linear method and kernel method. But, this paper just uses a linear method because it will compare sentiment and text, which only uses two variables.

3. **Random Forest**

   One of the strongest machine learning algorithms for classification is the Random Forests algorithm. Most prefer it because it delivers high predictive accuracy. It works by creating an ensemble of decision trees. Each is trained on a random subset of data and makes an independent prediction. These individual predictions are then combined through a majority voting mechanism to come up with the final classification. This ensemble strategy dramatically reduces the danger of overfitting, a major concern in decision tree models, by averaging out biases and variations from individual trees.Thus, they are popular in a variety of applications. [10]

Model Development: To develop the Naïve Bayes, SVM, and Random Forest models using appropriate libraries or frameworks, such as scikit-learn in the Python programming language.

Model Training: The models are trained using the training data, where features from the Twitter texts are extracted and fed into the model.

D. *Comparative Analysis and Evaluation*

The evaluation metrics that are suitable for this case are accuracy and F1-Score. Accuracy measures the portion of instances that are correctly labeled. F1-Score measures both precision and recall, recall measures how proficiently a model can identify actual positive cases, and precision measures the frequency models predict correctly the positive values. By comparing the accuracy and F1-Score results from Support Vector Machine, Naïve Bayes, and Random Forest, it can be concluded which one is more suitable for classification of Twitter data.

IV. Result And Discussion

The following are the findings from experiments conducted using the method that has been designed. This experiment has three methods to analyze sentiment. As shown in **(figure 2)**, the runtime for Naïve Bayes Classifier is 1 second. Followed by Random Forest, which takes 5 minutes and 2 seconds. Lastly, the runtime of SVM is 42 minutes and 49 seconds.

	Time
SVM	42 minutes 49 seconds
Random Forest	5 minutes 2 seconds
Naïve Bayes	1 second

Figure 2 : Model Training Times

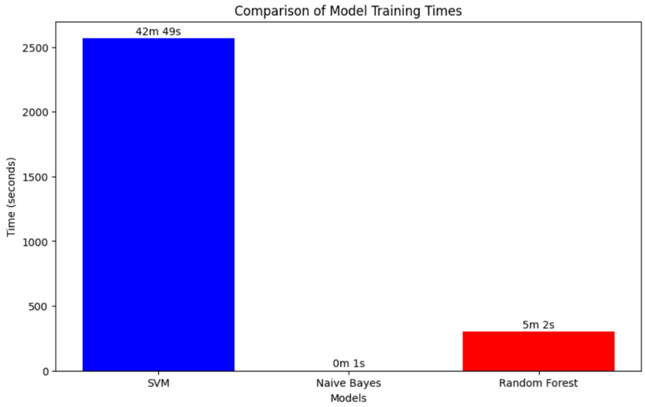

Figure 3 : Comparison of Model Training Times

The primary reason for the variations in runtime amongst the three sentiment analysis techniques—Naïve Bayes Classifier, Random Forest, and Support Vector Machine (SVM)—is their differing levels of algorithm complexity. Because Naïve Bayes relies on feature independence and a straightforward probabilistic technique, it is computationally cheap and incredibly quick. Random Forest takes longer since it creates several decision trees, each of which needs a great deal of data partitioning and analysis. Because SVM finds the ideal hyperplane for classification by solving a challenging quadratic optimization problem, a computationally demanding process that becomes even more difficult when dealing with huge datasets and non-linear kernels it is the slowest method.

As shown in (***figure 4***)**,** the accuracy and weighted average F1-score of SVM are 95.6% Followed closely by Random Forest with 95.5%. Lastly, Naïve Bayes has only 82.9% accuracy.

	Accuracy	F1-score
SVM	0.95596	0.95598
Random Forest	0.95496	0.95491
Naïve Bayes	0.82883	0.82828

Figure 4 : Classification Report SVM, Random Forest and Naïve Bayes

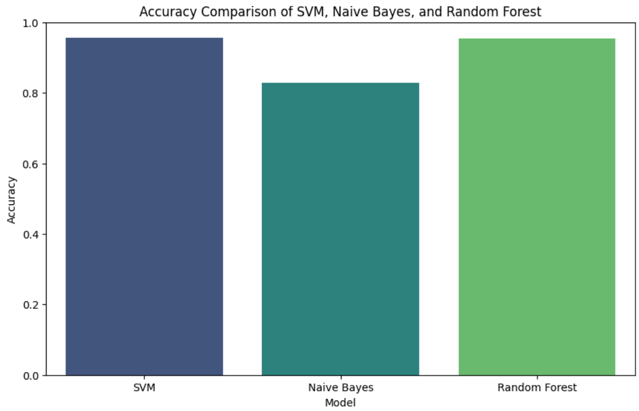

Figure 5 : Comparison of Model Accuracies

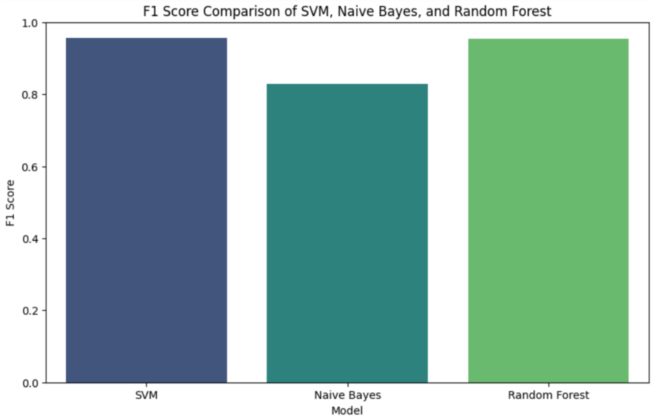

Figure 6 : Comparison of Model F1-scores

The SVM, Random Forest, and Naïve Bayes classifiers' notable variations in accuracy can be ascribed to a number of important aspects. SVMs work incredibly well with complicated datasets because they are strong in high-dimensional spaces and excel at identifying the best hyperplanes to divide data into different classes. Because Random Forests are ensemble approaches, they combine many decision trees to capture complicated feature interactions while also controlling overfitting and improving accuracy. Conversely, Naïve Bayes relies on the assumption of feature independence, which frequently fails to hold true in real-world data, making it less capable of capturing intricate correlations. Furthermore, substantial hyperparameter adjustment improves the performance of SVM and Random Forest models, while Naïve Bayes takes less tuning but provides less flexibility. The type of data also matters; SVM and Random Forest perform better on datasets with complex feature correlations and class imbalances. Class weighting and sampling are two strategies that these models can use to keep up performance. Comparing SVM and Random Forest to the more straightforward and heavily predicated Naïve Bayes classifier, their overall improved accuracy is indicative of their enhanced capacity to represent intricate patterns and relationships in data.

V. Conclusion

Three machine learning methods are used in this paper to analyze the sentiment of Twitter comments. This work concludes that, although Naïve Bayes has the quickest training time, it also has the lowest accuracy, which makes it less appropriate for intricate sentiment data analysis on large datasets such as Twitter comment data. However, SVM and Random Forest provide better accuracy and performance metrics; despite Random Forest's greater runtime, SVM slightly outperforms it in terms of overall performance. The findings suggest that while requiring more computation, SVM and Random Forest are great for accurate opinion categorization in scenarios like Twitter comment records where high precision as well as noisy data tolerance are needed.

References

[1] Statista, "Number of X (formerly Twitter) users worldwide from 2019 to 2024," Statista, 2023. [Online]. Available: https://www.statista.com/statistics/303681/twitter-users-worldwide/ [Accessed: March 4, 2024].

[2] 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 Proceedings of the 2019 8th International Conference on System Modeling and Advancement in Research Trends, SMART 2019, 2020. doi: 10.1109/SMART46866.2019.9117512.

[3] Styawati, Styawati, et al. "Comparison of Support Vector Machine and Naïve Bayes on Twitter Data Sentiment Analysis," Jurnal Informatika Politeknik Harapan Bersama, vol. 6, no. 1, 31 Jan. 2021, pp. 56-60, doi:10.30591/jpit.v6i1.3245.

[4] N. N.- IJSRP and undefined 2018, "Twitter sentiment analysis using support vector machine and K-NN classifiers," researchgate.netN NawIJSRP, 2018•researchgate.net, 2018, doi: 10.29322/IJSRP.8.10.2018.p8252.

[5] M. Ahmad, S. Aftab, I. A.-Int. J. Comput. Appl, and undefined 2017, "Sentiment Analysis of Tweets using SVM," researchgate.netM Ahmad, S Aftab, I AliInt. J. Comput. Appl, 2017•researchgate.net, vol. 177, no. 5, pp. 975–8887, 2017, doi: 10.5120/ijca2017915758.

[6] N. Bahrawi, "Sentiment Analysis Using Random Forest Algorithm-Online Social Media Based," *Journal of Information Technology and Its Utilization*, vol. 2, no. 2, pp. 29–33, Dec. 2019, doi: 10.30818/JITU.2.2.2695.

[7] Y. Rianto and A. Y. Kuntoro, "Prediction of Netizen Tweets Using Random Forest, Decision Tree, Naïve Bayes, and Ensemble Algorithm," *Sinkron : jurnal dan penelitian teknik informatika*, vol. 5, no. 1, pp. 58–71, Sep. 2020, doi: 10.33395/SINKRON.V5I1.10565.

[8] A. Talita, O. Nataza, Z. R.-J. of physics: conference, and undefined 2021, "Naïve bayes classifier and particle swarm optimization feature selection method for classifying intrusion detection system dataset," iopscience.iop.org, doi: 10.1088/1742-6596/1752/1/012021.

[9] Q. J. Liu, L. H. Jing, and L. M. Wang, "The Development and Application of Support Vector Machine," Journal of Physics: Conference Series, vol. 1748, no. 5, p. 052006, Jan. 2021, doi: 10.1088/1742-6596/1748/5/052006.

[10] J. Hatwell, M. Gaber, R. A.-A. I. Review, and undefined 2020, "CHIRPS: Explaining random forest classification," Springer, vol. 53, no. 8, pp. 5747–5788, Dec. 123AD, doi: 10.1007/s10462-020-09833-6.