Sentiment Analysis of YouTube User Comments using SVM and Naïve Bayes Method

Louis Garrick Setiadi
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
louis.setiadi@binus.ac.id

Franklin Pinehas Liauw Romamti
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
franklin.romamti@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*— The rapid growth of YouTube has generated vast amounts of user comments, requiring advanced methods to analyze sentiment. This paper compares the performance of Naive Bayes and Support Vector Machines (SVM) in classifying YouTube comments into positive and negative sentiments using Natural Language Processing (NLP) techniques. Our results show that SVM achieves higher accuracy than Naive Bayes. This study highlights the effectiveness of machine learning in sentiment analysis and discusses challenges such as informal language and sarcasm.**

Keywords—YouTube, Sentiment Analysis, Machine Learning, Naïve-Bayes, SVM

I. Introduction

Within the current digitalization time, we are undoubtedly familiar with different social media stages serving as settings for worldwide communication and interaction. Among these stages, YouTube stands out as an unmistakable illustration. Nearby other social media stages such as Instagram, Facebook, Twitter, and numerous more, YouTube positions moment all-inclusive regarding client predominance. According to reports by We Are Social and Hootsuite, YouTube reached 2.51 billion clients in January 2023. Besides its work as a stage for uploading video substance, YouTube also encourages community-building, the trade of suppositions through comment segments, and serves as a medium for getting input. Moreover, YouTube serves as a stage for trade wanders and salary era, provoking clients from around the globe to compete for brilliance as substance makers. One noteworthy perspective of YouTube's social media scene is its comment areas, where clients express their suppositions and lock in in discourses. This angle presents an interesting subject for estimation investigation.

The exponential development of online substance renders manual examination illogical. Thus, the development of computerized opinion investigation is fundamental. Apart from machine learning, deep learning techniques are also widely utilized in sentiment analysis [10]. Leveraging normal dialect handling (NLP) strategies, the estimation classification of YouTube comments can be categorized into positive, negative, and neutral sentiments [8, 9]. Through machine learning calculations and NLP strategies, analysts can observe designs in winning patterns and opinions communicated through YouTube comments, yielding positive client benefits [9].

This paper will use and compare two machine learning-based approaches, namely Support Vector Machines and Naive Bayes. Our essential objective in conducting this assumption examination is to supply a comprehensive understanding of estimations related to different themes, recordings, and channels on YouTube. This ponder points to evaluating the adequacy and exactness of opinion examination in analyzing client assumptions passed on in client comments. Inside this setting, a few challenges emerge in opinion investigation, including casual dialect utilization, mockery, and lexical uncertainty. Thus, these challenges oblige the exactness of opinion investigation comes about. In this manner, the taking after questions serve as the premise for the issue definition: (1) How effective are the Naive Bayes and Support Vector Machine methods in classifying YouTube comments?; (2) What are the factors that can influence the failure of the algorithm model's accuracy in classifying YouTube comments?; (3) Which of the two methods can provide more accurate results and how big is the difference in accuracy?

We will accumulate information from different YouTube comments traversing different themes and recordings to support our investigation. The information sources will contain a few YouTube client comments labeled concurring with pre-existing assumption categories. With these information sources, we trust to set up an establishment to assess the adequacy and precision of opinion classification strategies inside the setting of YouTube comments.

II. Related Work

In [1], sentiment analysis was carried out on YouTube comments to measure the level of attitude of users towards aspects of the video that they explained in a sentence. In research conducted by researchers Ritika Singh and Ayushka Tiwari, they compared several methods, including Naive Bayes and SVM. From the results of this research, based on several metrics used to test these methods, such as Macro scores, Micro scores, and F1 scores, it was found that SVM has better performance than other classifiers with an average accuracy level of 88%. Besides that, Naive Bayes has the second-best performance after SVM, with an average accuracy rate of 87%. This research shows that SVM is proven to be more accurate than Naive Bayes in measuring the attitude level of YouTube users in commenting on a video.

Based on previous research in general, our study aims to participate in testing the level of effectiveness and reliability of the Naive Bayes and SVM methods in conducting sentiment analysis on YouTube comments. Even though Naive Bayes and SVM are popular methods used by many researchers in conducting sentiment analysis, there are still often some failures in the analysis. Therefore, our research also aims to explore and analyze several factors that can cause analysis failure and why this happens.

in [2], there is also research that analyzes the Naive Bayes method to find heuristic solutions in perfecting the Multinomial Naive Bayes (MNB) classification algorithm for sentiment analysis on film review datasets based on positive or negative sentiment labels. This research was conducted to address systemic issues with the MNB model, improving classification accuracy, and demonstrating the effectiveness of the MNB classifier through in-depth experiments with various text classification datasets to highlight significant improvements in terms of sentiment analysis using the MNB algorithm.

in [3], a study conducted sentiment analysis on several opinion data on political candidates through comments and tweets using SVM and ANN. The results obtained show that the performance of SVM is quite good in categorizing text compared to ANN, this shows the ability of SVM to handle high-dimensional feature spaces effectively.

In addition, In [4], there is research that develops a cross-domain sentiment analysis (CDSA) model on Indonesian language YouTube comment data to deal with stop words and slang words that are unique to the Indonesian dialect. In this research, it was found that the best classifier model for CDSA in Indonesian was the Extra Tree method with an accuracy level of 91.91%, while for Naive Bayes and SVM, it gave results with accuracy levels of 79.54% and 89.35%. Here it can also be seen that although SVM is not the best method in terms of CDSA in Indonesian, it still produces results that tend to be better than Naive Bayes.

In [5], a study using the Multinomial Naive Bayes (MNB) and Logistic Regression (LR) algorithms with the Count Vectorizer and TF-IDF methods on the Hausa language Tweets dataset sourced from the BBC Hausa Twitter account, this research shows a significant difference in performance between the two the algorithm. MNB using the TF-IDF method gets accuracy results of 80% and MNB using the CountVectorizer method gets accuracy results of 80% as well. Meanwhile, Multinomial Logistic Regression with the TF-IDF method got an accuracy of 83% and Multinomial Logistic Regression with the CountVectorize method got an accuracy of 86%. Based on these results, the Logistic Regression algorithm shows a better level of classification accuracy in categorizing Hausa language text compared to MNB, although Naive Bayes shows better performance in terms of speed and memory usage for large datasets.

III. METHODOLOGY

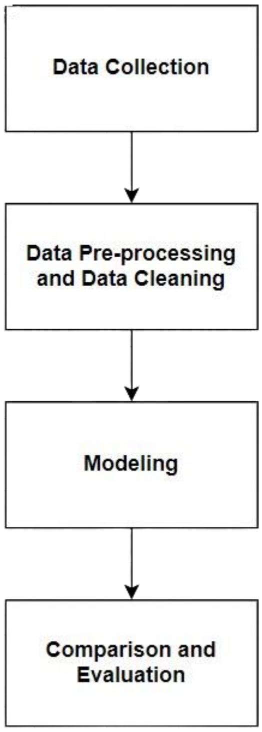

***Figure 1* : Methodology Process Diagram**

A. *Data Collection*

Our research will use quantitative methods in finding and analyzing the level of accuracy of each classifier. The sample dataset that we will use comes from Comments.csv which consists of 32,000 YouTube comment data with comment and sentiment variables (0 and 1) which will symbolize negative or positive sentiment. We will use this dataset as training data to train the level of accuracy of the algorithm method that we will use.

B. *Data Processing and Data Cleaning*

After the dataset is imported, data quality checks and data cleaning will be carried out to remove redundant and irrelevant data. Tokenization, removal of punctuation marks and symbols, removal of stop words, stemming, and lemmatization will be carried out to support performance during the text classification process.

C. *Modeling*

The Naive Bayes and Support Vector Machine algorithm approach will be the methods used to carry out this classification. Both are known as fairly reliable and most frequently used classification methods.

1. **Naïve Bayes**

The Naive Bayes algorithm is a popular classification method utilized in Sentiment Analysis to categorize text. It is a machine learning-based approach founded on Bayes' theorem, which provides the foundation for probability and statistics. The algorithm supposes that the features employed to make predictions are unrelated to each other [6]. In essence, the Naive Bayes algorithm employs a formula as such:

$$P(y|X) = P(X|y) \times P(y)/P(X) \quad (1)$$

In the classification context, X denotes the input feature vector, and y is the anticipated class label. During the prediction stage, the algorithm determines the class with the highest posterior probability $P(y|X)$ [6].

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

Support Vector Machine (SVM) is a machine learning algorithm often used in data classification and regression. The way SVM works is to find an optimal hyperplane that can separate two classes in feature space based on the given data samples [7]. There are several formulas for the SVM algorithm:

Linear Decision Function:

$$F(x) = w * x + b \quad (2)$$

Hyperplane:

$$W * x + b = 0 \quad (3)$$

The formulas above are the basic formulas for the SVM algorithm to find the best hyperplane that can separate two classes with a maximum margin [7].

Model Training: To train the Naive Bayes and Support Vector Machine algorithm models, 80% of the data from the dataset will be used for model training.

Model Testing: An analysis will be carried out on the performance of the model that has been trained. To test this analysis, 20% of the data from the dataset will be used to evaluate the model.

D. Comparison and Evaluation

Several evaluation metrics will be involved as indicators of comparison of the two methods. The precision metric is used to measure how often the model correctly predicts comments with a "positive" or "negative" sentiment. The Recall metric is used to measure how accurate the model is at analyzing all comments that are truly “positive” or “negative”. F1 Score is used to see the overall performance of the model in predicting both sentiments (between precision and recall metrics). The Support metric is used to see the number of comments that have a "positive" or "negative" sentiment. Thus, the final performance results will be calculated by the Accuracy metric and include the Macro Average and Weighted Average metrics as supporting metrics.

IV. RESULT AND DISCUSSION

Based on the methodology that we have previously described, at this stage we will model each method we use, namely Naive Bayes and SVM. Here, we will use the F1 Score, Accuracy, Recall, Precision, Macro Average, and Weighted Average metrics. To check whether comments are included in positive or negative sentiment, we use indicators 1 and 0, where if comment = 1 it will be included in positive comments and if comment = 0 it will be included in negative comments. Detailed results of metric calculations in Naive Bayes can be shown in (***Figure 2***) and the results of metric calculations in SVM can be shown in (***Figure 3***).

As shown in (***Figure 4***), After conducting training and testing data on positive (1) and negative (0) sentiment, we obtained a difference in accuracy level of 25% between SVM and Naive Bayes, namely with an accuracy percentage of 85% and 60%. The following are the results of experiments that have been carried out based on the methodology that has been designed:

	Precision	Recall	F1-score	Support
Negative	1.00	0.11	0.20	2859
Positive	0.58	1.00	0.74	3541
Macro Avg	0.79	0.56	0.47	6400
Weighted Avg	0.77	0.60	0.50	6400

Figure 2 : **Metric results from Naive Bayes**

	Precision	Recall	F1-score	Support
Negative	0.90	0.01	0.02	980
Positive	0.85	1.00	0.92	5420
Macro Avg	0.87	0.50	0.47	6400
Weighted Avg	0.86	0.85	0.78	6400

Figure 3 : **Metric results from Support Vector Machine**

	Accuracy
Naïve Bayes	0.60
Support Vector Machine	0.85

Figure 4 : **Accuracy comparison between Naive Bayes and Support Vector Machine**

Below is a graph showing the comparison of accuracy levels and F1 scores between Naive Bayes and Support Vector Machine:

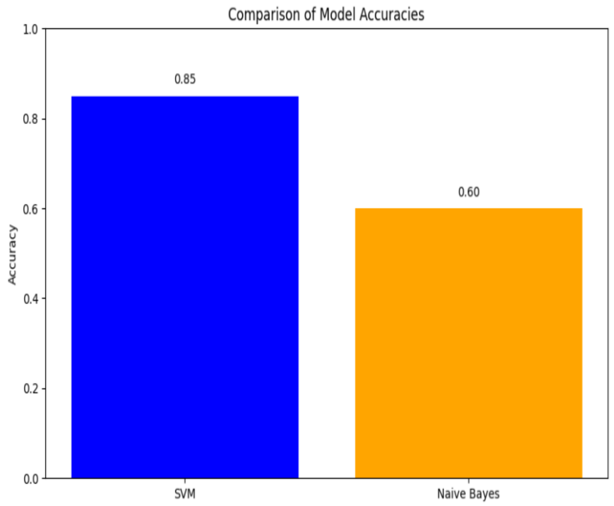

***Figure 5 :* Graph of accuracy level of Naive Bayes and SVM**

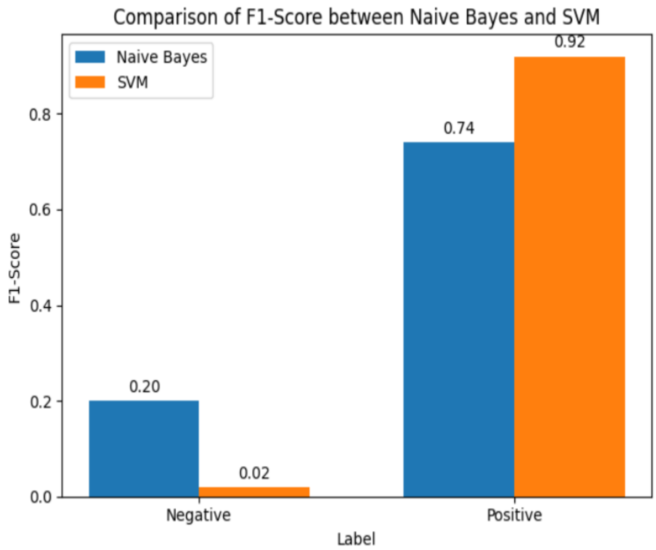

***Figure 6 :* Graph of F1 Score of Naive Bayes and SVM**

Based on the graph (***Figure 5***), shows that the Support Vector Machine provides a higher level of accuracy than Naive Bayes, namely 85% and 60%. Based on the analysis of the results of metric calculations in (***Figure 2***) and (***Figure 3***), Naive Bayes tends to predict negative sentiment more often than SVM, resulting in an accuracy level of Naive Bayes' F1 Score for negative sentiment of 20% and 74% for positive sentiment, while for SVM gives F1 Score results for negative sentiment of 2% and 92% for positive sentiment (***Figure 6***). To further clarify the final results, the Macro Average and Weighted Average metrics are used to calculate the average of the Precision, Recall, and F1 Score values. Thus, providing final results that show that the average performance score of SVM is higher than Naive Bayes. Despite the success of SVM and Naive Bayes in classifying YouTube comment data, it was still detected that there was a failure rate in data accuracy during the classification process. This may be caused by some system failures at the pre-processing and data-cleaning stages that do not detect ambiguous words, sarcasm, and abbreviations.

V. CONCLUSION

Based on the experiments we have carried out, we have obtained sentiment analysis results from a collection of YouTube comment data from various topics which show an accuracy level using the SVM method of 85% and Naive Bayes of 60%. Both methods demonstrate their respective effectiveness in classifying YouTube comments. Although, there are still visible data inaccuracies due to several data factors that are difficult to detect by the system, however, we have succeeded in comparing the level of effectiveness of the two methods and obtained results that prove that the SVM method has a higher level of accuracy than the Naive Bayes method.

REFERENCES

[1] R. Singh and A. Tiwari, "YOUTUBE COMMENTS SENTIMENT ANALYSIS," *International Journal of Scientific Research in Engineering and Management (IJSREM)*, vol. 5, no. 5, p. 11, May 2021.

[2] M. Abbas, K. A. Memon, A. A. Jamali, S. Memon, and A. Ahmed, "Security Multinomial Naive Bayed Classification Model for Sentiment Analysis," *International Journal of Computer Science and Network (IJCSNS)*, vol. 19 No 13, p. 5, Mar. 2019, [Online]. Available: https://www.researchgate.net/profile/Anees-Ahmed-5/publication/334451164_Multinomial_Naive_Bayes_Classification_Model_for_Sentiment_Analysis/links/5e227e8d92851cafc38c813c/Multinomial-Naive-Bayes-Classification-Model-for-Sentiment-Analysis.pdf

[3] G. Patil, V. Galande, V. Kekan, and K. Dange, "Sentiment Analysis Using Support Vector Machine," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2, no. 1, p. 5, Jan. 2014, [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=61e4c1aa3e344b0797e5187d677017a3004afdb5

[4] A. S. Aribowo, H. Basiron, F. Noor Abd Yusof, and S. Khomsah, "Informatics Cross-domain Sentiment Analysis Model on Indonesian YouTube Comment," *International Journal of Advances in Intelligent*, vol. 7 No 1, p. 14, Mar. 2021, [Online]. Available: https://pdfs.semanticscholar.org/277a/7f0ab34874b4dea02570e39c9f048b3db1c2.pdf

[5] M. Sani, A. Ahmad, and H. S Abdulazeez, "Sentiment Analysis of Hausa Language Tweet Using Machine Learning Approach," *Journal of Research in Applied Mathematics*, vol. VOL 8, no. 9, p. 16, 2022.

[6] R. Rahayu, "ALGORITMA NAÏVE BAYES," *ALGORITMA NAÏVE BAYES*, p. 7, Dec. 2023.

[7] L. Frank, "Support Vector Machines (SVMs)," https://www.researchgate.net/publication/380881348_Support_Vector_Machines_SVMs, p. 12, May 2024.

[8] D. Shah and M. Parekh, "From YouTube Comments to Insights: a sentiment analysis of opinions on productivity tools," *International Journal for Research in Applied Science & Engineering Technology*, vol. 11, no. 8, p. 8, Aug. 2023.

[9] F. Olaoye and K. Potter, "Natural language processing and sentiment analysis," *File:///C:/Users/Frank/Downloads/NaturalLanguageProcessingandSentimentAnalysis.pdf*, p. 14, Mar. 2024.

[10] P. Kumar Rangarjan *et al.*, "The Social Media Sentiment Analysis Framework: Deep learning for sentiment analysis on social media," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 3, p. 12, Jun. 2024.