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

Author 1
Louis Garrick Setiadi
2602050601

Author 2
Franklin Pinehas Liauw Romamti
2602095650

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

Introduction

YouTube, a prominent social media platform with 2.51 billion users as of January 2023, facilitates community-building and opinion exchange through its comment sections, making it an interesting subject for sentiment analysis. Given the vast amount of content, manual analysis is impractical, necessitating automated sentiment analysis using machine learning and deep learning techniques. This study aims to compare the effectiveness of Support Vector Machines (SVM) and Naive Bayes in classifying YouTube comments into positive, negative, and neutral sentiments. The research will address challenges such as casual language, sarcasm, and lexical ambiguity, focusing on evaluating the accuracy and efficacy of these methods. Data will be collected from various YouTube comments to establish a foundation for assessing the performance of these sentiment classification techniques.

Literature Review

Previous research has compared the effectiveness of Naive Bayes (NB) and Support Vector Machines (SVM) for sentiment analysis of YouTube comments, showing that SVM generally outperforms NB in accuracy, with SVM achieving 88% and NB 87%. Ritika Singh and Ayushka Tiwari's study found SVM to be more accurate than NB, while other studies have explored enhancing the Multinomial Naive Bayes (MNB) algorithm for better performance in sentiment analysis. SVM has also demonstrated strong performance in political sentiment analysis and handling high-dimensional feature spaces. Cross-domain sentiment analysis on Indonesian YouTube comments revealed that the Extra Tree method was most accurate, but SVM still outperformed NB. Additionally, research on Hausa language tweets showed that Logistic Regression outperformed MNB in accuracy, though MNB was faster and more memory-efficient. This body of research aims to further test and understand the effectiveness and reliability of NB and SVM in sentiment analysis, investigating factors contributing to potential analysis failures.

Methodology

Data Collection:
We will use a dataset of 32,000 YouTube comments with sentiment labels (0 for negative, 1 for positive) from Comments.csv to train and test our classifiers.

Data Processing and Cleaning:
Data cleaning will include removing redundant data, tokenization, punctuation and stop word removal, stemming, and lemmatization.

Modeling:

- Naive Bayes: Utilizes Bayes' theorem to classify text based on probability.
- Support Vector Machine (SVM): Finds an optimal hyperplane to separate classes in feature space.

Comparison and Evaluation:
We will use precision, recall, F1 score, support, accuracy, macro average, and weighted average metrics to compare and evaluate the performance of the two methods.

Result and Discussions

Metric results from Naive Bayes

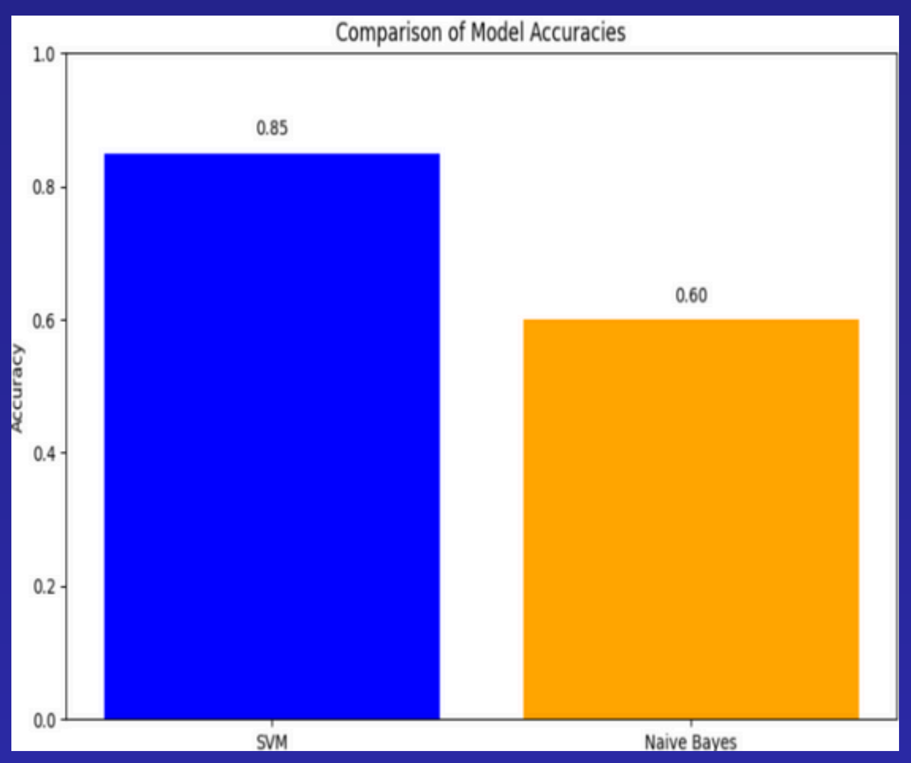
	Precision	Recall	F1-score	Support
Negative	1.00	0.11	0.20	2859
Positive	0.58	1.00	0.74	3541
Macro Avg	0.79	0.56	0.47	6400
Weighted Avg	0.77	0.60	0.50	6400

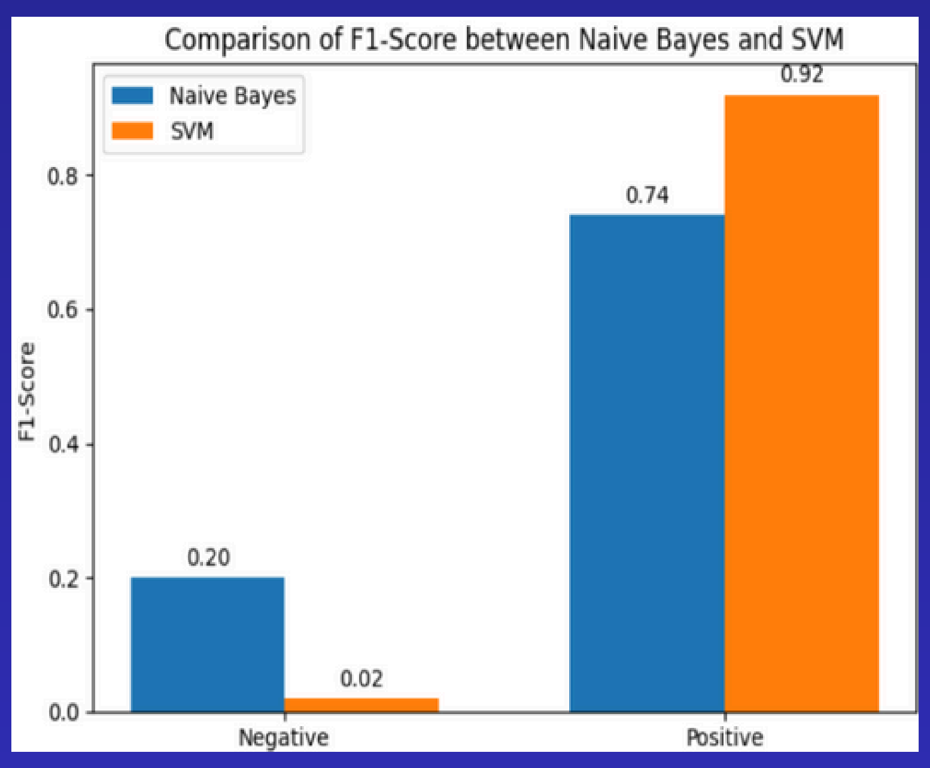
Metric results from Support Vector Machine

	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

Accuracy comparison between Naive Bayes and Support Vector Machine

	Accuracy
Naïve Bayes	0.60
Support Vector Machine	0.85





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.