Journal of Student Research Exploration

https://shmpublisher.com/index.php/josre
p-ISSN 2964-8246| e-ISSN 2964-1691

Comparison of naïve bayes and support vector machine methods for jkt48 music video comment classification

Alif Abdul Aziz[1], **Rofik**[2]
[1, 2]Department of Computer Science, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received July 9, 2024
Revised January 28, 2025
Accepted February 8, 2025

Keywords:

Classification
Naive bayes
Sentiment analysis
Support vector machine
Youtube comment

ABSTRACT

The research was conducted to discuss the classification of comments on music video JKT48 "Magic Hour" in YouTube using method Naive Bayes Classifier (NBC) and Support Vector Machine (SVM). YouTube monitors viewer emotion by adjective comments Adjectives are the descriptive powers of human communication we use to help personify how different types, i.e. different "personalities" flavors and depths reflect artistic expressions The place where interactivity meets with digital marketing signifying a shared contribution to music lore in this work, we study the comparison of The Support Vector Machine (SVM) and Naive Bayes Classifier in terms of Accuracy, Precision & Recall. This Project includes data pre-processing, collecting the data by YouTube API and build classification models which involves Support Vector Machine and Naive Bayes Classifier. SVM displayed more stable performance than NBC, showing consistent results across different data split ratios. SVM achieved its highest accuracy of 93.42% at an 80:20 ratio, with precision and recall rates reaching 92.57% and 93.42%, respectively.

1. Introduction

The provision of universal service and access to information and communication technologies is a key national objective in many countries, often enshrined in

[1] ***Corresponding Author:***
Alif Abdul Aziz,
Information System Study Program,
Universitas Negeri Semarang,
Kampus Sekaran, Gunung Pati, Semarang City, Indonesia .
Email: alifabdulaziz24@students.unnes.ac.id
DOI: https://doi.org/10.52465/josre.v3i1.389

legislation governing the sector [1]. YouTube, as one of the largest video-based platforms in the world, allows users to leave comments expressing their opinions, criticisms, or appreciation of the content they view. YouTube has become an international platform for formal and informal learning [2]. YouTube has had a significant impact on information and communication technology, especially seen in music videos. It is also utilised by various demographics, such as secondary education students in Kazakhstan, to improve their language skills [3]. Additionally, YouTube serves as an important marketing communication platform, influencing brand preference and appeal among Millennials in developing countries such as Romania and South Africa [4]. Besides its role in music video sharing, YouTube significantly impacts a wide range of content creators, from beauty vloggers to cultural bridge builders, who influence everyday information practices and foster global communities [5].

In the context of community management and sentiment analysis, prior research looks at several YouTube-specific tasks. s is widely used as the Naive Bayes classifier for this purpose. Sentiment analysis [6], [7] is one of its uses and when applied to certain subset like cyberbullying detection [8], spam comments discovery [9], [10], accuracy achieved are often considerably high. For instance, in one experiment on Naive Bayes Classifier 92.78% accuracy has been achieved for spam comment detection and 84.11 % of sentiment analysis [Stemlis : Recipes Bargaining commence]. When the Naive Bayes Classifier was hybridized with other base classifiers as Support Vector Machine, Logistic Regression, k-nearest neighbors (KNN), Decision Tree and Random Forest it increased even more the accuracy for comment classification between positive or negative levels achieving 94.62% with a combination of different ones based on Naïve Bayes [6]. We have also seen the use of Support Vector Machine for classification of YouTube video comments. This algorithm is able to process complex and heterogeneous data [11], which has lead it to being increasingly exploited by both scientific researches as well practical applications for sentiment analysis or content filtering. As such, a Support Vector Machine has been used for both comment polarity classification [10] and spam commentary recognition towards COVID-19 news videos [12], as well as public sentiment analysis of vaccination policies on YouTube [13]. It is in fact shown by research that Support Vector Machine commonly outperforms Naive Bayes Classifier for new spam comments on YouTube videos, from the perspective of accuracy [14], [15].

Even though many ways exist to investigate YouTube comments, a literature gap on evaluating the methods with each other in one context still remains. Most literature will typically investigate a single kind of method or comment type in isolation without performing direct comparisons with other methods on the same dataset. Moreover, while former research almost completely had not to count words in comments on music video from JKT48 who abundantly produced diverse comment. Hence, further evaluation of the performance of various category

methods in YouTube comment classification is required particularly on music videos.

This research attempts to bridge this gap as it compares the performance of Naive Bayes Classifier and Support Vector Machine algorithms in classifying 1000 YouTube comments on a JKT48 music video titled "Magic Hour". The goal of this research is to compare the accuracy, speed and generalizability of these two different approaches as a part of comment sentiment analysis. Confidently saying that this research will be an important and of high valued contribution on social media sentiment analysis field by generating the necessary information to digital content managers how they can take action in order, user experience towards platform change for better.

2. Method

This section describes the method used to compare and investigate which is more effective, Naive Bayes or Support Vector Machine(SVM) algorithm in text classification of comments on music video "Magic Hour" by JKT48. The steps are include data collection using YouTube API preprocessing like cleaning and transformation classification model - Naive Bayes, SVM evaluation of metrics including accuracy, precision-recall The procedure consists of following steps, as shown in the Figure 1 below:

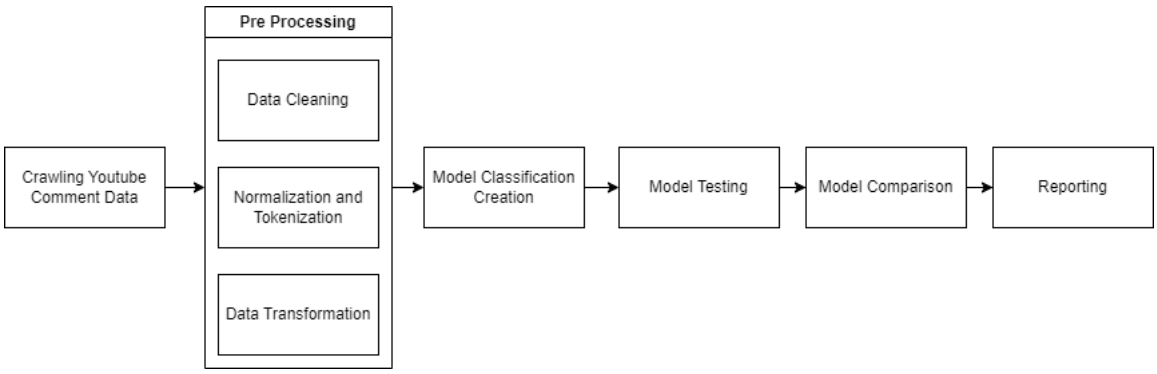

Figure 1. Flowchart

Data Collection

This study uses data from YouTube, which specifically retrieves comments related to the JKT48 music video "Magic Hour. Comment data was collected through the YouTube API, which allows for authorized access to some of comment metadata (e.g., usernames comments and when these were added) ordered by popularity. Once the data was generated, it was further imported into a Python development environment using Pandas library for in depth analysis. 1000 Comments were collected to cover all kind of comments.

Preprocessing

Preprocessing of data means putting the table in a form that is easier to analyze - closer to something we could use for Analysis [16]. This process involves stages cleaning (data cleansing), normalization or standardization data and transformation [17] aligning data in the format. These are important steps to prepare the data for what the system needs. Prepropcess preprocessing used in this study:

This preprocessing includes data cleansing processes such as removing duplicated, irrelevant or empty comments from the dataset and dealing with missing values by imputing new data based on existing features in this way to clean out all unclean elements.

Text is normalized to remove capitalization and extraneous characters and stopwords like 'and', 'which' or the. The text data undergoes tokenization to split it into words, after which each word might be further organically processed commonly with stemming being applied where needed (reduction of a stem word).

TF-IDF (Term Frequency-Inverse Document Frequency) and Word2Vec are used to change the textual data into numbers which is called as Data Transformation. TF-IDF gives weights to words in relation to the frequency of that word across documents and also, corpus whereas Word2Vec tries to establish semantic relationship between similar meaning words. These methods make it feasible to be processed by classification algorithms and therefore, improve textual sentiment analyses of remarks.

Naive Bayes Classification

Naive Bayes is a classification method that applies Bayes' theorem to assess the probability of a document belonging to a specific category, taking into account the likelihood of observed features within the document [18]. This approach utilizes Bayes' theory, which derives the posterior class distribution from the prior class distribution and the conditional probability distribution (CPD) of features. Naive Bayes assumes feature independence when classifying examples, although this assumption may not always hold true. Nonetheless, in practical applications, Naive Bayes often achieves comparable or superior performance compared to more complex learning methods [19]. Bayes' theorem in its general form is expressed as follows [20]:

The expression of Bayes' Theorem for one proof and one hypothesis can be seen in (1):

$$P(H|E) = \frac{P(E|H).P(H)}{P(E)} \quad (1)$$

with:

- P(H|E) represents the probability of hypothesis H given evidence E.
- P(E|H) denotes the probability of evidence E given hypothesis H.

- P(H) stands for the probability of hypothesis H independent of any evidence.
- P(E) indicates the probability of evidence E.

Bayes' theorem for one proof and multiple hypotheses can be formulated as in (2):

$$P(Hi|E) = \frac{P(E|Hi).P(Hi)}{\sum_{k=1}^{n} P(E|Hk).P(Hk)} \quad (2)$$

with:

- P(Hi|E) signifies the probability of hypothesis Hi being true given evidence E.
- P(E|Hi) represents the probability of evidence E given hypothesis Hi is true.
- P(Hi) denotes the probability of hypothesis Hi irrespective of any evidence.
- n represents the total number of hypotheses.

Support Vector Machine Classification

Support Vector Machine (SVM) is a classification algorithm used to classify datasets, whether linear or non-linear [21]. At its core, SVM aims to find a hyperplane that effectively separates data points into distinct classes. The computations of SVM involve utilizing the following formula:

Data point:

$$x_i = \{x1, x2, \ldots, xn\} \in Rn \quad (3)$$

Data class:

$$y_i \in \{-1, +1\} \quad (4)$$

The representation of data and classes can be seen in (5):

$$\{(x_i, y_i)\}_{i=1}^{N} \quad (5)$$

Maximize the function using the formula (6) :

$$Ld = \sum_{i=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \; terms: 0 \le \alpha_i \le C \; and \; \sum_{i=1}^{N} \alpha_i y_i = 0 \quad (6)$$

Calculating w and b values by using formula (7):

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \quad b = -\frac{1}{2}(w.x^{+} + w.x^{-}) \quad (7)$$

Classification decision function sign(f(x)), as in (8):

$$f(x) = w.x + b \quad or \quad f(x) = \sum_{i=1}^{m} \alpha_i y_i K(x, x_j) + b \quad (8)$$

Description:

- N : Number of data points
- n : Number of features/dimensions of the data
- C : Constant value used in the SVM algorithm
- M : Number of support vectors/data points for which $\alpha i > 0$
- K(x,xi): Kernel function used in SVM
- Ld : Lagrange multiplier duality
- αi : Weight value associated with each data point

3. Results and Discussion

The collected data undergoes preprocessing steps which include automated and manual removal of missing values. Subsequently, filtering, tokenization, stemming, and automatic labeling processes are applied. In this study, we evaluate the performance of two classification models, Naive Bayes (NB) and Support Vector Machine (SVM), to categorize YouTube comments into positive, negative, and neutral sentiments. The evaluation is conducted using various test dataset sizes (test_size), specifically 0.1, 0.2, 0.3, 0.7, 0.8, and 0.9. We assess the models based on accuracy, precision, recall, and F1 score for each test dataset size. Detailed calculation results are presented in Table 1.

Table 1. Test results

Method	Classification				Accuracy %	Precision (%)	Recall (%)
	Ratio (Train & Test)	Positive	Negative	Neutral			
NBC	10:90	53	7	280	82.35	67.82	82.35
	20:80	47	7	248	82.12	67.44	82.12
	30:70	42	6	216	81.82	66.94	81.82
	70:30	13	2	99	88.60	88.17	88.60
	80:20	9	1	66	89.47	89.30	89.47
	90:10	4	1	33	89.47	87.98	89.47
SVM	10:90	53	7	280	82.94	83.81	82.94
	20:80	47	7	248	83.11	83.68	83.11
	30:70	42	6	216	82.95	83.62	82.95
	70:30	13	2	99	89.47	88.86	89.47
	80:20	**9**	**1**	**66**	**93.42**	**92.57**	**93.42**
	90:10	4	1	33	92.11	90.13	92.11

Performance of Naive Bayes Classifier (NBC) Method

Indeed, a consistent aspect of the data splitting experiments was that the performance of our Naive Bayes classifier (NBC) method fluctuated with changes in training and test set sizes. The NBC tests found the below specific results.

Accuracy, precision, and recall of naive bayes classifier

Tests were conducted with data split ratios of 10:90, 20:80, 30:70, 70:30, 80:20, and 90:10. The results indicate that NBC achieved its highest accuracy of 88.60% at a 70:30 ratio, accompanied by precision and recall rates of 88.17% and 88.60%, respectively. However, there was a notable decline at the 90:10 ratio, where NBC achieved accuracy, precision, and recall of 89.47%, 87.98%, and 89.47%, respectively.

variations in naive bayes classifier performance

Across different sizes of training and test data, NBC demonstrated significant performance fluctuations. Despite achieving high accuracy in certain data split ratios, NBC exhibited greater variability compared to SVM.

Performance of Support Vector Machine (SVM) Method

Next, the performance of a vey popular method in sentiment classification applied to YouTube comment dataset called Support Vector Machine (SVM) is analyzed over varying training and test data size configurations.

Accuracy, precision, and recall of support vector machine

SVM displayed more stable performance than NBC, showing consistent results across different data split ratios. SVM achieved its highest accuracy of 93.42% at an 80:20 ratio, with precision and recall rates reaching 92.57% and 93.42%, respectively. Even at a 90:10 ratio, SVM maintained robust performance with an accuracy of 92.11%, precision of 90.13%, and recall of 92.11%.

Consistency of support vector machine performance

The results underscore the stability of SVM's performance compared to NBC. Despite minor fluctuations in performance at certain data split ratios, SVM consistently delivered high accuracy and reliable precision and recall metrics.

The Implications of the Findings

This observation puts emphasis on the importance of not only accuracy when choosing a classification method (as follows from traditional crossvalidation), but also in terms of stability over different test scenarios. SVM is the best choice for sentiment analysis of YouTube comments where performance needs to be consistent and always reliable. That said, NBC also presented competitive results under specific data split settings.

4. Conclusion

Using the JKT48 music video "Magic Hour" as a case study, this study compared the performance of the Naive Bayes classifier and the Support Vector Machine (SVM) for sentiment classification in YouTube comments. SVM proved to be more effective on smaller datasets with less clear separation, while Naive Bayes tended to overfit on larger datasets, but performed well with higher data sharing ratios. This shows that sentiment analysis methods need to balance accuracy and performance robustness. Future research should explore continuous data coverage, factors such as ensembles and hyperparameter optimisation, and comparisons with other classification methods to further refine our understanding of sentiment analysis in this context.

REFERENCES

[1] S. Sarkar, "The Role of Information and Communication Technology (ICT) in Higher Education for the 21st Century," *Sci. Probe*, vol. 1, Jan. 2012.

[2] A. Shoufan and F. Mohamed, "YouTube and Education: A Scoping Review," *IEEE Access*, vol. 10, pp. 125576–125599, 2022, doi: 10.1109/ACCESS.2022.3225419.

[3] A. Toleuzhan, G. Sarzhanova, S. Romanenko, E. Uteubayeva, and G. Karbozova, "The Educational Use of YouTube Videos in Communication Fluency Development in English: Digital Learning and Oral Skills in Secondary Education," *Int. J. Educ. Math. Sci. Technol.*, vol. 11, no. 1, pp. 198–221, Nov. 2022, doi: 10.46328/ijemst.2983.

[4] M. Kyaw Sein, "Information and Communication Technology for Development (ICT4D)," *Inf. Soc.*, vol. 35, no. 2, pp. 107–108, Mar. 2019, doi: 10.1080/01972243.2019.1568715.

[5] R. Duffett, D.-M. Petroșanu, I.-C. Negricea, and T. Edu, "Effect of YouTube Marketing Communication on Converting Brand Liking into Preference among Millennials Regarding Brands in General and Sustainable Offers in Particular. Evidence from South Africa and Romania," *Sustainability*, vol. 11, no. 3, p. 604, Jan. 2019, doi: 10.3390/su11030604.

[6] S. Mulyani and R. Novita, "Implementation Of The Naive Bayes Classifier Algorithm For Classification Of Community Sentiment About Depression On Youtube," *J. Tek. Inform.*, vol. 3, no. 5, pp. 1355–1361, Oct. 2022, doi: 10.20884/1.jutif.2022.3.5.374.

[7] M. Subramanian, V. E. Sathishkumar, K. Shanmugavadivel, P. Deva, S. Haris, and J. Cho, "Detecting homophobic and transphobic texts from youtube comments using machine learning models," *Appl. Comput. Eng.*, vol. 6, no. 1, pp. 952–961, Jun. 2023, doi: 10.54254/2755-2721/6/20230958.

[8] Ahlida Nikmatul H, Didih Rizki C, and Christian S.K. Aditya, "Classification of Bullying Comments on YouTube Streamer Comment Sections Using Naïve Bayes Classification," *J. Syst. Eng. Inf. Technol.*, vol. 2, no. 1, pp. 25–28, Mar. 2023, doi: 10.29207/joseit.v2i1.5016.

[9] H. Valpadasu, P. Chakri, P. Harshitha, and P. Tarun, "Machine Learning based Spam Comments Detection on YouTube," in *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, May 2023, pp. 1234–1239. doi: 10.1109/ICICCS56967.2023.10142608.

[10] D. A. Musleh *et al.*, "Arabic Sentiment Analysis of YouTube Comments: NLP-Based Machine Learning Approaches for Content Evaluation," *Big Data Cogn. Comput.*, vol. 7, no. 3, p. 127, Jul. 2023, doi: 10.3390/bdcc7030127.

[11] S. U. Ahsaan, H. Kaur, A. K. Mourya, and S. Naaz, "A Hybrid Support Vector Machine Algorithm for Big Data Heterogeneity Using Machine Learning," *Symmetry (Basel).*, vol. 14,

no. 11, p. 2344, Nov. 2022, doi: 10.3390/sym14112344.
[12] S. Sudianto, J. A. A. Masheli, N. Nugroho, R. W. Ananda Rumpoko, and Z. Akhmad, "Comparison of Support Vector Machines and K-Nearest Neighbor Algorithm Analysis of Spam Comments on Youtube Covid Omicron," *J. Tek. Inform.*, vol. 15, no. 2, pp. 110–118, Dec. 2022, doi: 10.15408/jti.v15i2.24996.
[13] A. Wijayanto, "Analisis Sentimen Komentar Youtube Mengenai Vaksin Covid-19 Menggunakan Support Vector Machine," *J. PILAR Teknol. J. Ilm. Ilmu Ilmu Tek.*, vol. 7, no. 1, pp. 24–31, Jun. 2022, doi: 10.33319/piltek.v7i1.118.
[14] A. Sinhal and M. Maheshwari, "An Extensive Review on Contemporary Analysis of Comment Filtration of YouTube Videos Using Machine Learning Techniques," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 12, no. 9, pp. 130–143, Sep. 2022, doi: 10.46338/ijetae0922_14.
[15] M. B. Puneeth and V. Ramakrishnan, "The Mechanism of Spam Comment Detection Using Count Vectorizer and Naive Bayes Machine Learning Algorithms in Python," *ECS Trans.*, vol. 107, no. 1, pp. 13417–13428, Apr. 2022, doi: 10.1149/10701.13417ecst.
[16] D. Maulina and M. Corry Andhara, "Perbandingan Pre-Processing Opini Netizen Terhadap RUU PKS Menggunakan Algoritma Naive Bayes Classifier," *Smart Comp Jurnalnya Orang Pint. Komput.*, vol. 12, no. 1, Jan. 2023, doi: 10.30591/smartcomp.v12i1.4610.
[17] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Glob. Transitions Proc.*, vol. 3, no. 1, pp. 91–99, Jun. 2022, doi: 10.1016/j.gltp.2022.04.020.
[18] D. D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," 1998, pp. 4–15. doi: 10.1007/BFb0026666.
[19] I. Rish, "An empirical study of the naive Bayes classifier," in *IIJCAI 2001 workshop on empirical methods in artificial intelligence*, 2001, pp. 41–46.
[20] Y. Junaedi, B. N. Sari, and A. S. Y. Irawan, "Sistem Pakar Untuk Diagnosis Hama Pada Tanaman Jambu Air Menggunakan Metode Theorema Bayes," *J. Ilm. Inform.*, vol. 5, no. 2, pp. 168–178, Dec. 2020, doi: 10.35316/jimi.v5i2.960.
[21] D. Prajarin, "Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Kulit," *INFORMAL Informatics J.*, vol. 1, no. 3, pp. 137–141, 2016.