KHATULISTIWA SMART : JOURNAL OF ARTIFICIAL INTELLIGENCE
e-ISSN : XXXX-XXXX
p-ISSN : XXXX-XXXX

Khatulistiwa Smart : Journal of Artificial Intelligence
Vol. 1 No. 1 (2021) : 35-40
Available online at : https://journal.literasikhatulistiwa.org/index.php/smart

Sentiment Analysis of Hate Speech Using SVM Method

Hayatul Kamalia [1]
[1] Universitas Nurul jadid, Probolinggo
hayatulkamalia76@gmail.com[1]

Doi
Received: 28 November 2024 *Accepted: 21 Desember 2024* *Published: 31 Desember 2024*

Abstract This research analyses the sentiment of hate speech on social media using the Support Vector Machine (SVM) method. The Indonesian dataset from Kaggle is processed through text normalisation, filtering, and stemming to ensure the data is suitable for use in machine learning models. The SVM model was compared with Naive Bayes and Random Forest. Results showed SVM excelled with 75.40% accuracy, compared to Naive Bayes (67.34%) and Random Forest (46.64%). Performance evaluation is done with a confusion matrix that measures accuracy, precision, recall, and F1-score. The advantage of SVM lies in its ability to find optimal decision boundaries in a multidimensional feature space, making it more effective in handling complex interactions between features compared to Naive Bayes and Random Forest. The findings show that SVM is more effective for the classification of hate speech on social media. This research contributes to the development of automated monitoring systems that are more accurate and efficient in detecting and classifying hate speech content, thus improving countermeasures on social media platforms.

Keywords *Hate Speech; Sentiment Analysis; SVM; Machine Learning; NLP*

1. INTRODUCTION

In recent years, advances in intelligent technology and the application of machine learning techniques have revolutionised various fields, including electronic media such as Twitter, Instagram, Facebook, and YouTube. These platforms allow information to spread quickly and widely (Noola & Basavaraju, 2022). Social media has experienced rapid development and is one of the sectors that has benefited the most from this technological advancement (Han et al., 2020). However, in addition to its benefits, social media also presents new challenges, one of which is the spread of hate speech (Hana et al., 2020). In Indonesia, the freedom of speech guaranteed by the government is often misused by some people to spread hate speech, which harms minority groups based on ethnicity, race, and religion (Rajeena P. P et al., 2023).

In today's digital era, access to information is very fast, making it difficult to distinguish between true and false information. This condition has led to an increase in hate speech cases in society (Liang, 2021). Handling hate speech cases is not easy and requires a lot of human resources. Therefore, there is a need for a predictive model that is able to detect hate speech automatically to reduce this burden (Ulfah & Anam, 2020). The main purpose of hate speech classification is to analyse spoken words and determine their likelihood of being hate speech, neutral, or not hate speech (Husada & Paramita, 2021). The dataset used in this research is obtained from public data on Kaggle in Indonesian, which is then translated into English for data processing (Oryza Habibie Rahman et al., 2021), (Tineges et al., 2020).

Several studies have been conducted to detect hate speech using datasets from social media such as Twitter, but most of them focus on the classification of hate speech in English text (Shannaq et al., 2022).

Research (Liang, 2021) focuses on the classification of hate speech on social media in Indonesian, using three machine learning algorithms: SVM, XGBoost, and Neural Network. The results showed that Neural Network with RMSProp optimiser gave the best results with 82.9% accuracy, 82% precision, 82% recall, and 82% F1-score. This research focuses on the problem of hate speech classification on social media, especially on the Twitter platform, where hate speech is widespread and requires effective methods for automatic classification. The main objective of this research is to develop and evaluate a classification model that can separate hate speech from non-hatred speech. In addition, this research compares the performance of several machine learning algorithms, namely Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and Neural Network, in performing such classification. The measurement matrix used to evaluate model performance includes accuracy, precision, recall, and F1-score. The evaluation results show that the Neural Network with the RMSProp optimiser provides the best results with an accuracy of 82.9%, precision of 82%, recall of 82%, and F1-score of 82%. Recent research (Darwis et al., 2020) This research highlights the challenges of classifying diverse and unstructured public opinion on social media, such as Twitter, specifically related to views on the Corruption Eradication Commission (KPK). The goal is to classify public sentiment towards the KPK based on data from Twitter, so that the KPK can understand public opinion and improve its performance based on that feedback. The method used is Support Vector Machine (SVM). The test results show that SVM has an accuracy of 82%, using a confusion matrix to display the prediction of positive, negative, and neutral classes.

2. RESEARCH METHODS

This section describes the research methods used to build and evaluate the performance of the model. Information is required such as the method chosen to obtain the data set, data preparation techniques, data analysis techniques, etc. as shown in Figure 1.

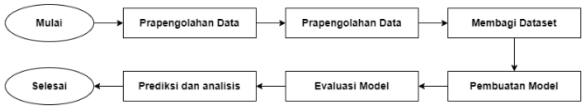

Figure 1. Process flow diagram

2.1. Dataset

This research uses a dataset derived from kaggle data which consists of 7368 Three classes are Positive, Neutral, and Negative. This dataset consists of 3029 Positive. 1401 neutral, 2285 Negative following samples taken from the dataset.

Table 1. Sampel Dataset

No	Text	Label
1	tangkap ahok rakyat bersamaf pi indonesia akan hancur tanpa ulama	Negatif
2	status dukung gerakan matikan lilin untuk ahok gubernur sebut akun facebok dihack	Negatif
3	kisah seorang nenek mencuri singkong karena kelaparan dan hakim menangis sat menjatuhkan vonis	Netral
4	hanya ada satu dekan di satu perguruan tingi	Netral
5	saya pikir rasanya rata-rata . tidak terlalu istimewa dibanding dengan sate yang lain yang saya pernah makan di bandung	Positif

2.2. Data preprocessing

The data preprocessing stage is an important step to ensure that the raw text can be processed by machine learning models. These steps include text normalisation, where all text characters are converted to lowercase, punctuation marks are removed, and numbers are removed to unify the text format to facilitate further processing. Next, filtering is performed using the Natural Language Toolkit (NLTK) package to remove stopwords that do not provide significant information for analysis. After that, stemming was performed using the Sastrawi library, which converts words into their base form, presented in table 2. All of these preprocessing steps are combined in one text processing pipeline to ensure that each text is processed through all stages in the correct order, resulting in texts that are ready for further analysis. Once the texts are processed, the Indonesian texts are translated to English using the TextBlob library, examples of some of the translated ones are presented in table 3. which is necessary as the sentiment analysis in this study is done with a library that better supports the English language. Finally, sentiment analysis was performed using TextBlob, which provides tools to determine whether the sentiment of the translated text is positive, negative, or neutral. Furthermore, the dataset is divided into training data and testing data with a ratio of 70:30, where 70% of the data is used to train the model and 30% of the data is used to test the model.

Table 2. stemming menggunakan sastrawi

Text	Label
status dukung gerakan matikan lilin untuk ahok…	status dukung gera mati lilin untuk ahok guber
aki amien rais bilang prabowo dengan bung karn…	aki amien rais bilang prabowo dengan bung karn...
belajar dari negara tetanga vietnam terbukti k...	ajar dari negara tetanga vietnam bukti korupsi...

Table 3. Translate to english

Text	Label
status dukung gerakan matikan lilin untuk ahok…	support status due to the death of Ahok's cand...
aki amien rais bilang prabowo dengan bung karn…	aki amien rais says prabowo karno karno one tr...
belajar dari negara tetanga vietnam terbukti k...	teach neighboring countries, Vietnam, evidence...

2.3. Modelling and Model Evaluation

In this stage, model building, model evaluation, and prediction and analysis of prediction results using the Support Vector Machine (SVM) method with a linear kernel are carried out. The first step is to visualise the distribution of comp_score labels in the dataset using a pie chart, which gives an idea of the proportion of each label. Next, the text data is cleaned from empty values and converted into numerical representation using TF-IDF Vectorizer. The cleaned text data is then divided into training data and testing data with a ratio of 70:30 using the train_test_split function, ensuring a balanced distribution of labels in both datasets. After that, the SVM model is trained using the training data. The trained model is then used to predict the labels on the test data. The prediction accuracy is calculated, and a classification report that includes metrics such as precision, recall, and f1-score is presented in Table 4 to evaluate the model performance. In addition, the confusion matrix presented in Figure 2 was created to visualise the model's performance in predicting correct and incorrect labels. The evaluation results show the accuracy level of the model as well as other performance metrics, providing an overall picture of how well the model can generalise to new data. Visualisations and tables of such evaluation results will be shown to support this analysis.

Table 4. laporan klasifikasi precision, recall, dan f1-score

	precision	recall	f1-score	support
neg	0.74	0.75	0.75	710
neu	0.63	0.58	0.60	420
pos	0.82	0.83	0.82	939
accuracy			0.75	2069

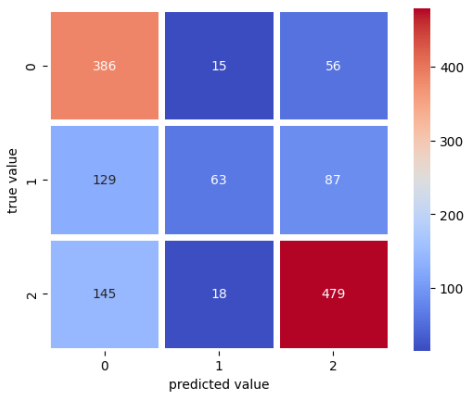

Figure 1. Process flow diagram

3. RESULTS AND DISCUSSION

In this study, an evaluation of the Support Vector Machine (SVM) algorithm for sentiment analysis of hate speech was conducted. The analysis results show that the SVM model provides the best performance with an accuracy of 75.40%. To compare the effectiveness of SVM, this research also includes experiments with two other machine learning algorithms, namely Naive Bayes and Random Forest. Naive Bayes obtained an accuracy of 67.34%, while Random Forest only achieved an accuracy of 46.64%. The advantage of SVM in handling the complexity of text data lies in its ability to find the optimal decision boundary in a multidimensional feature space, which enables better detection of subtile patterns in text. In contrast, Naive Bayes, which is based on the assumption of feature independence, is less effective in capturing complex interactions between features. Random Forest, despite being a frequently used ensemble algorithm, did not perform as expected in this case, possibly due to its inability to handle highly imbalanced data or closely related features. These findings confirm that SVM is a more effective choice for social media hate speech classification compared to the methods tested in comparison. These results make an important contribution to the development of automated monitoring systems that more accurately and efficiently handle hate speech content on social media platforms.

4. CONCLUSIONS

This research focuses on sentiment analysis of hate speech using the Support Vector Machine (SVM) method. The evaluation results show that SVM provides the best performance with an accuracy of 75.40%. Comparison with two other algorithms, Naive Bayes and Random Forest, shows that SVM excels in handling text data complexity. Naive Bayes achieved an accuracy of 67.34%, while Random Forest only 46.64%. The advantage of SVM lies in its ability to find the optimal decision boundary in a multidimensional feature space, enabling better detection of subtle patterns in text compared to Naive Bayes and Random Forest. Naive Bayes, assuming feature independence, is ineffective in capturing complex interactions between features, while Random Forest is inadequate in dealing with highly imbalanced data. The results of this study confirm that SVM is a more effective method for the classification of hate speech on social media. These findings make an important contribution to the development of automated monitoring systems that are more accurate and efficient in handling hate speech content on social media platforms. By using SVM, the monitoring system can better detect and classify hate speech, improving the effectiveness of countering hate speech on social media.

5. REFERENCES

Darwis, D., Pratiwi, E. S., & Pasaribu, A. F. O. (2020). Penerapan Algoritma Svm Untuk Analisis Sentimen Pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia. *Edutic - Scientific Journal of Informatics Education*, *7*(1), 1–11. https://doi.org/10.21107/edutic.v7i1.8779

Han, K. X., Chien, W., Chiu, C. C., & Cheng, Y. T. (2020). Application of support vector machine (SVM) in the sentiment analysis of twitter dataset. *Applied Sciences (Switzerland)*, *10*(3). https://doi.org/10.3390/app10031125

Hana, K. M., Adiwijaya, Al Faraby, S., & Bramantoro, A. (2020). Multi-label Classification of

Indonesian Hate Speech on Twitter Using Support Vector Machines. *2020 International Conference on Data Science and Its Applications, ICoDSA 2020*, *August 2020*. https://doi.org/10.1109/ICoDSA50139.2020.9212992

Husada, H. C., & Paramita, A. S. (2021). Analisis Sentimen Pada Maskapai Penerbangan di Platform Twitter Menggunakan Algoritma Support Vector Machine (SVM). *Teknika*, *10*(1), 18–26. https://doi.org/10.34148/teknika.v10i1.311

Liang, S. (2021). Comparative Analysis of SVM, XGBoost and Neural Network on Hate Speech Classification. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *5*(5), 896–903. https://doi.org/10.29207/resti.v5i5.3506

Noola, D. A., & Basavaraju, D. R. (2022). Corn leaf image classification based on machine learning techniques for accurate leaf disease detection. *International Journal of Electrical and Computer Engineering*, *12*(3), 2509–2516. https://doi.org/10.11591/ijece.v12i3.pp2509-2516

Oryza Habibie Rahman, Gunawan Abdillah, & Agus Komarudin. (2021). Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *5*(1), 17–23. https://doi.org/10.29207/resti.v5i1.2700

Rajeena P. P, F., S. U, A., Moustafa, M. A., & Ali, M. A. S. (2023). Detecting Plant Disease in Corn Leaf Using EfficientNet Architecture—An Analytical Approach. *Electronics (Switzerland)*, *12*(8). https://doi.org/10.3390/electronics12081938

Shannaq, F., Hammo, B., Faris, H., & Castillo-Valdivieso, P. A. (2022). Offensive Language Detection in Arabic Social Networks Using Evolutionary-Based Classifiers Learned From Fine-Tuned Embeddings. *IEEE Access*, *10*(June), 75018–75039. https://doi.org/10.1109/ACCESS.2022.3190960

Tineges, R., Triayudi, A., & Sholihati, I. D. (2020). Analisis Sentimen Terhadap Layanan Indihome Berdasarkan Twitter Dengan Metode Klasifikasi Support Vector Machine (SVM). *Jurnal Media Informatika Budidarma*, *4*(3), 650. https://doi.org/10.30865/mib.v4i3.2181

Ulfah, A. N., & Anam, M. K. (2020). Analisis Sentimen Hate Speech Pada Portal Berita Online Menggunakan Support Vector Machine (SVM). *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, *7*(1), 1–10. https://doi.org/10.35957/jatisi.v7i1.196