Future Computer Science Journal (FCSJ)
Vol. 1, No. 1, June 2023: xxx-xxx
http://asasijournal.id/index.php/fcsj
http://doi.org/10.xxxxx/fcsj.xxxx.xxx

Stacking Ensemble Learning Technique for Sentiment Analysis of Hate Speech on Twitter

Yopi Julia Nurriski[1*], Triani Femilia Agustina[2], Jumanto[3]
[1,2,3]Computer Science Department, Faculty of Mathematics and Natural Science, Universitas Negeri Semarang, Indonesia

Abstract

Hate speech on social media, especially on the Twitter platform, is an increasing problem. The spread of negative, demeaning, or threatening messages can have a negative impact on individuals and society at large. Previous research focuses more on improving accuracy in general without looking at the f1-score on tweets categorized as negative. Therefore, this research aims to address these issues by improving the accuracy and f1-score of the overall sentiment analysis of hate speech on Twitter. Through the use Stacking Ensemble Learning technique, It was found that the combination of models consisting of Support Vector Machine (SVM), Decision Tree, and Random Forest was able to provide the most accurate results with an accuracy rate of 96.16% and F1 Score of 96.13%. This research demonstrates the potential of using these techniques to effectively and efficiently identify and address hate speech on Twitter, and contributes to creating a more positive and safe online environment.

Keywords:
Hate speech;
Sentiment analysis;
Stacking Ensemble Learning;
Support Vector Machine;
Decision Tree;

Article History:
Received: June 19, 2023
Revised: June 20, 2023
Accepted: June 20, 2023
Published: June 20, 2023

Corresponding Author:
Yopi Julia Nurriski
Computer Science Department, Universitas Negeri Semarang, Indonesia
Email:
yopijulianurriski@students.unnes.ac.id

INTRODUCTION

Hate speech is generally defined as a form of expression intended to demean, belittle, or harm individuals or groups based on certain characteristics, such as race, ethnicity, national origin, religion, sexual orientation, gender identity, or other factors [1]. Nowadays, with the development of social media, hate speech is increasingly easy to find, especially on the Twitter platform. Twitter, as one of the largest social media platforms in the world, makes it easy for users to share information quickly and easily. However, this convenience also provides a gap for irresponsible users to disseminate hate speech.

Sentiment analysis techniques are used to perform opinion classification. However, it is debatable whether this opinion classification technique can be used to classify hate speech [2]. This is because hate speech has its own characteristics and is not always related to the opinion or sentiment expressed. Therefore, several studies have been conducted to develop specialized techniques that can distinguish between hate speech and general opinions on social media. Hate speech has been of deep concern, as it can lead to negative impacts such as the spread of hatred, discrimination, and social polarization. In this context, sentiment analysis becomes an important tool to identify and classify hate speech on platforms such as Twitter.

Madhu, et al. [3] have developed sentiment analysis techniques by developing a pipeline using a fine-tuned SentBERT paired with LSTM as a classifier. This pipeline achieved a macro F1 score of 0.892 on the ICHCL test dataset. Another study was also conducted by Aouchiche, et al. [4] by developing random forest and autoencoder models as deep learning models combined with random forest models. The proposed model achieved 94% and 63% accuracy. Min, et al. [5] also conducted similar research that developed the Emotion Correlated Hate Speech DetectOR (EHSor) method and succeeded in significantly improving HSD performance compared to existing HSD methods. However, in this study, it was found that the ability of EHSor to predict samples categorized as 'hate' weakened after the super-hate detector feature was removed and there was an imbalance in the data distribution of the evaluated dataset. As a result, the accuracy and precision metrics showed an increase while the other metrics showed a decrease. Therefore, there is a need to develop a better performing hate speech sentiment analysis model that is able to overcome the imbalance in data distribution in the dataset so that it can have a positive impact.

Based on the existing problems, a sentiment analysis technique is proposed, namely "Stacking Ensemble Learning Technique for Sentiment Analysis of Hate Speech on Twitter". Decision Tree is a method that is widely used in classifying data [6]. Decision tree is a model that uses a tree structure to represent a series of decisions based on dataset features. Each node in the tree represents a decision based on certain feature values, and each branch of that node represents a possible decision result that is easy to understand [7]. In addition, the Support Vector Machine algorithm is effective in handling data with many features and is used in various fields such as face recognition, handwriting, and text classification [8]. The use of Random Forest plays a role in overcoming the problem of overfitting that occurs in the decision tree during the training process [9]. Currently, there are many popular classification models, one of which is stacking ensemble learning [11]. Stacking ensemble learning is a method that combines several different machine learning models or algorithms to improve prediction accuracy [12], [13]. In improving accuracy, Meta-learner can be used to improve the performance of machine learning models or algorithms by learning patterns or characteristics of the model [14], [15], [16]. Therefore, in this research, the Stacking Ensemble Learning method will be used to combine Decision Tree, Support Vector Machine, and Random Forest models to improve the accuracy of sentiment analysis of hate speech on Twitter to learn the patterns or characteristics of these models.

METHOD

In this research, the analysis of hate speech on twitter uses several stages, namely collecting data, preprocessing data, sentiment analysis, data visualization, stacking ensemble classifier, and evaluation. The workflow of this research is shown in Figure 1.

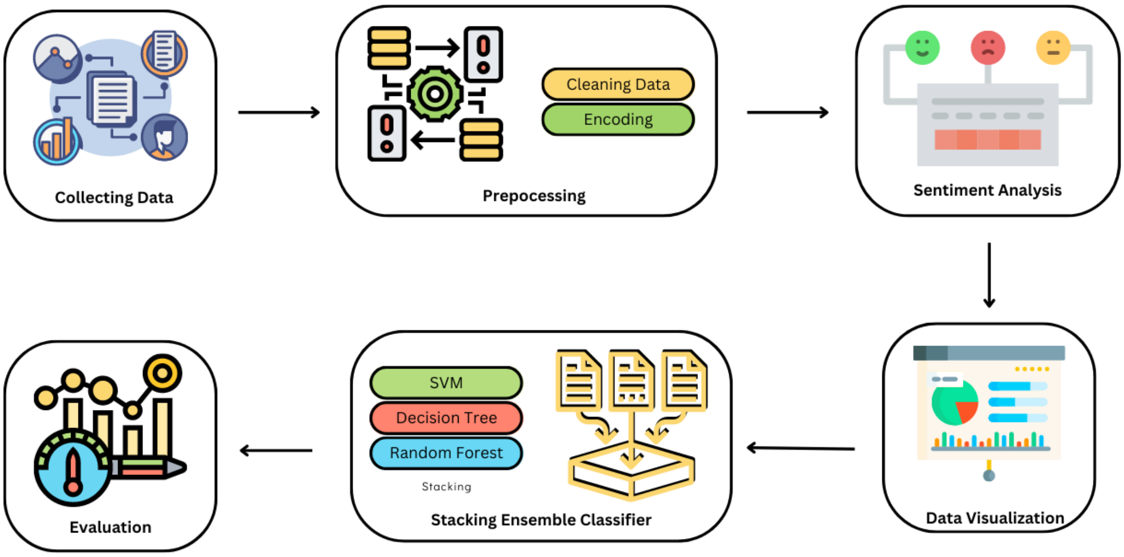

Figure 1. Workflow of Hate Speech Sentiment Analysis

2.1 Data Description

The data used in this research uses the Tweeter Hate Speech Analysis dataset which can be accessed on kaggle. This dataset is the main data source containing tweets that have been categorized as hate speech or not. The dataset used is 31962 data with three features, namely id, label, and tweet shown in Table 1.

Table 1. Example of Hate Speech Dataset Example of Hate Speech Sentiment Analysis Dataset on Twitter

id	label	tweet
1	o	@user when a father is dysfunctional and is s...
2	o	@user @user thanks for #lyft credit i can't us...
3	o	@user @user thanks for #lyft credit i can't us...
4	o	@user @user thanks for #lyft credit i can't us...
5	o	factsguide: society now #motivation

In the dataset to be used, there are still columns and many characters that are not needed so that preprocessing must be done first before the data is analyzed further [17]. The data to be analyzed then goes through the process of extracting relevant features, such as tweet text, hate speech category labels, and other attributes needed. However, in this research, relabeling is done to get labels that are suitable for sentiment analysis, namely normal, positive, and negative. By collecting data from Twitter Hate Speech Analysis, this research has a representative database to perform sentiment analysis of hate speech on Twitter using stacking ensemble learning techniques that combine Decision Tree, Support Vector Machine, and Random Forest algorithms.

2.2 ***Preprocessing***

The data preprocessing stage in this research is very important to prepare the data that will be used in sentiment analysis of hate speech on Twitter. First, the tweet text from the Tweeter Hate Speech Analysis dataset is processed to count NaN-valued data and then deleted. Second, tokenization is performed to break the text into separate word units so that it can be processed by a computer. Next, unnecessary features are removed from the dataset using the drop_features(features,data) function. Special characters, links, and other characters that can interfere with the analysis process are also removed with the function in the process_tweet(tweet) class which is then converted to lowercase. The clean tweet data is then stored in a new column with the name processed_tweet in the previous dataset table as shown in Table 2.

Table 2. Tweets Dataset after Cleaning the processed_tweets Column

id	label	tweet	processed_tweets
1	o	@user when a father is dysfunctional and is s...	when a father is dysfunctional and is so selfi...
2	o	@user @user thanks for #lyft credit i can't us...	thanks for lyft credit i can t use cause they ..
3	o	@user @user thanks for #lyft credit i can't us...	bihday your majesty
4	o	@user @user thanks for #lyft credit i can't us...	model i love u take with u all the time in ur
5	*o*	factsguide: society now #motivation	factsguide society now motivation

The next step is to remove the id, label, and tweet columns using the dataset.drop(columns = ['id', 'label', 'tweet']) function because these columns are not needed in sentiment analysis. The new table resulting from the column deletion is then stored in the data table shown in Table 3. This preprocessing stage aims to clean and prepare the twheet text to be ready for use in further sentiment analysis. With careful data preprocessing, tweet data from te Tweeter Hate Speech Analysis dataset will produce better analysis results.

Table 3. Tweet Dataset Ready for Analysis

	processed_tweets
1	when a father is dysfunctional and is so selfi...
2	thanks for lyft credit i can t use cause they ..
3	bihday your majesty
4	model i love u take with u all the time in ur
5	factsguide society now motivation

2.3 Sentiment Analysis

The sentiment analysis stage is a stage to extract and determine the sentiment or emotional attitude of the tweet text dataset [18]. In this research, tweet text sentiment analysis uses the TextBlob library in Python. There are three main functions implemented, namely getSubjectivity(processed_tweets), getPolarity(processed_tweets) and getAnalysis(score). The getSubjectivity(processed_tweets) function is a TextBlob sentiment object. This function is used to measure the subjectivity or objectivity of the text of a tweet using the subjectivity method. In addition, the getPolarity(processed_tweets) function is used to calculate polarity, which indicates the degree of positivity, negativity, or neutrality of the text. This function also uses the sentiment.polarity method of the TextBlob object. Meanwhile, the getAnalysis(score) function is used to return the sentiment classification based on the given score. If the score is 0, then the result is 'Neutral'. If the score is less than 0, then the result is 'Negative'. If the score is more than 0, then the result is 'Positive'. The sample results of the three main functions are presented in Table 4.

Table 4. Sample Sentiment Analysis Results from Tweets

	processed_tweets	Subjectivity	Polarity	Analysis
1	when a father is dysfunctional and is so selfi...	1.0	-0.5000	Negative
2	thanks for lyft credit i can t use cause they ...	0.2	0.2000	Positive
3	bihday your majesty	0.0	0.0000	Neutral
4	model i love u take with u all the time in ur	0.6	0.9766	Positive
5	factsguide society now motivation	0.0	0.0000	Neutral

2.4 Data Visualization

At this stage, the number of tweets based on the sentiment analysis results is calculated to visualize the dataset. The purpose of this data visualization is to show the percentage or number of tweets categorized as positive, negative, or neutral. This will help understand the proportion of sentiment present in the dataset. The pie chart shows that there are 11,149 tweets categorized as neutral, 16,118 tweets categorized as positive, and 4,696 tweets categorized as negative.

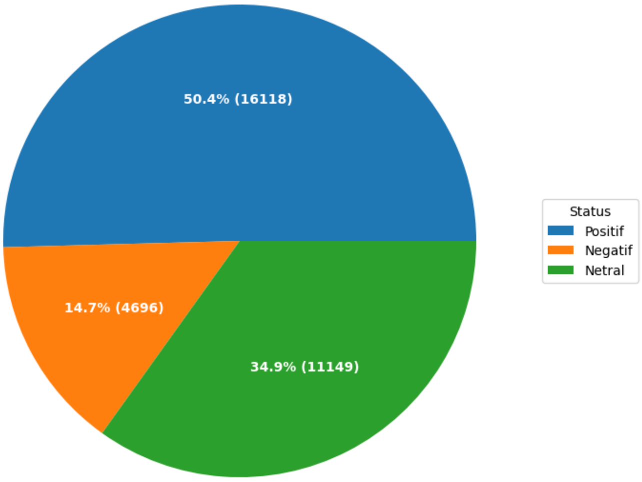

Figure 2. Circle Diagram of Sentiment Analysis Percentage

In addition to presenting data visualization in the form of diagrams, sentiment analysis is also presented in the form of wordclouds presented in Figure 3. for normal tweets, Figure 4. for positive tweets, and Figure 5. for negative tweets. By using an attractive visual representation such as wordcloud, the goal is to facilitate understanding and observation of the most dominant or common words in the neutral, positive, and negative sentiment categories. The wordcloud visualization will give attention to words that have a high frequency of occurrence in the text.

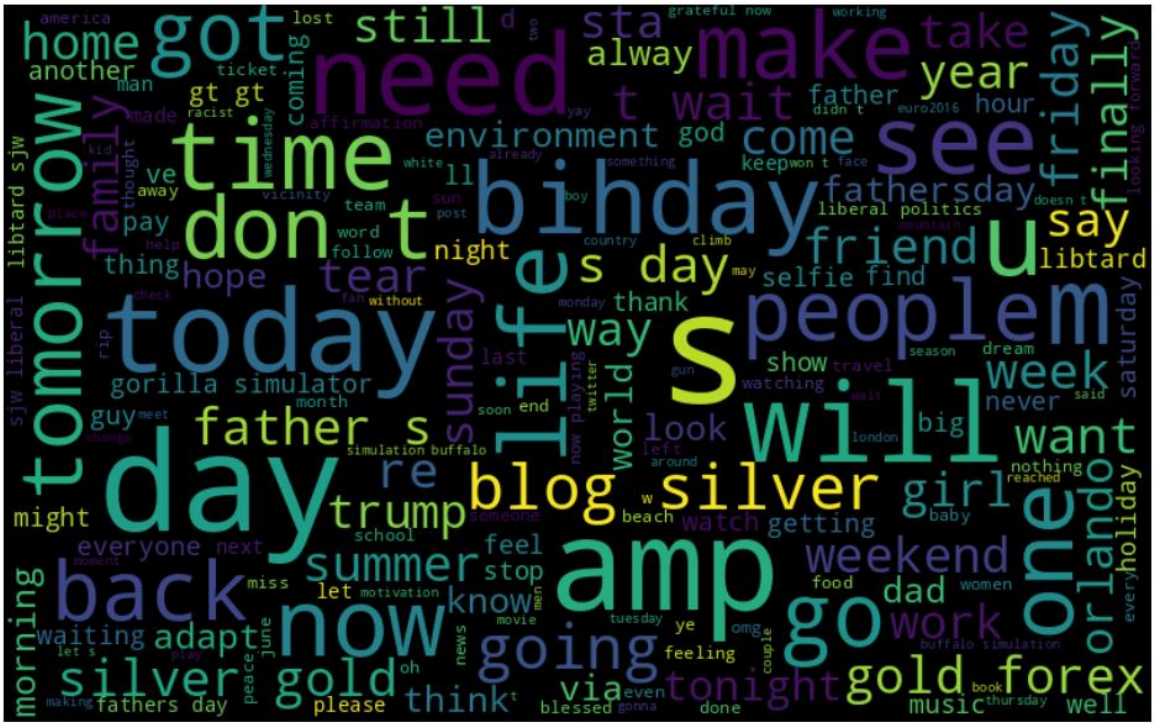

Figure 3. Neutral Tweet

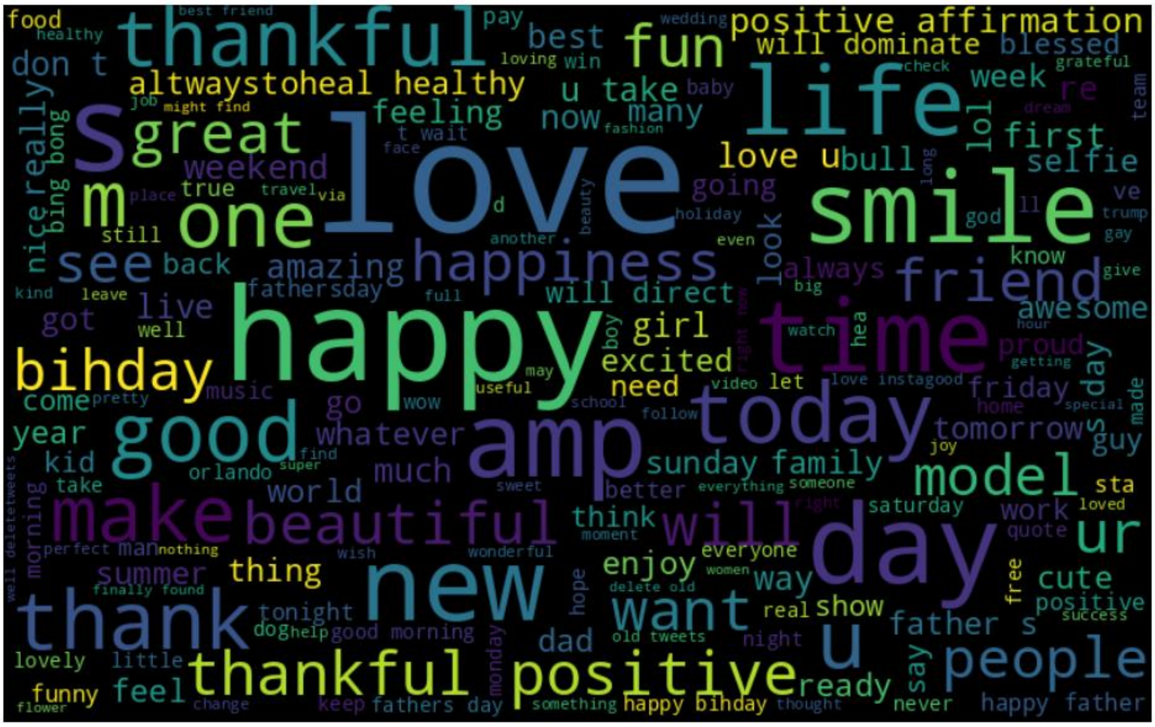

Figure 4. Positive Tweet

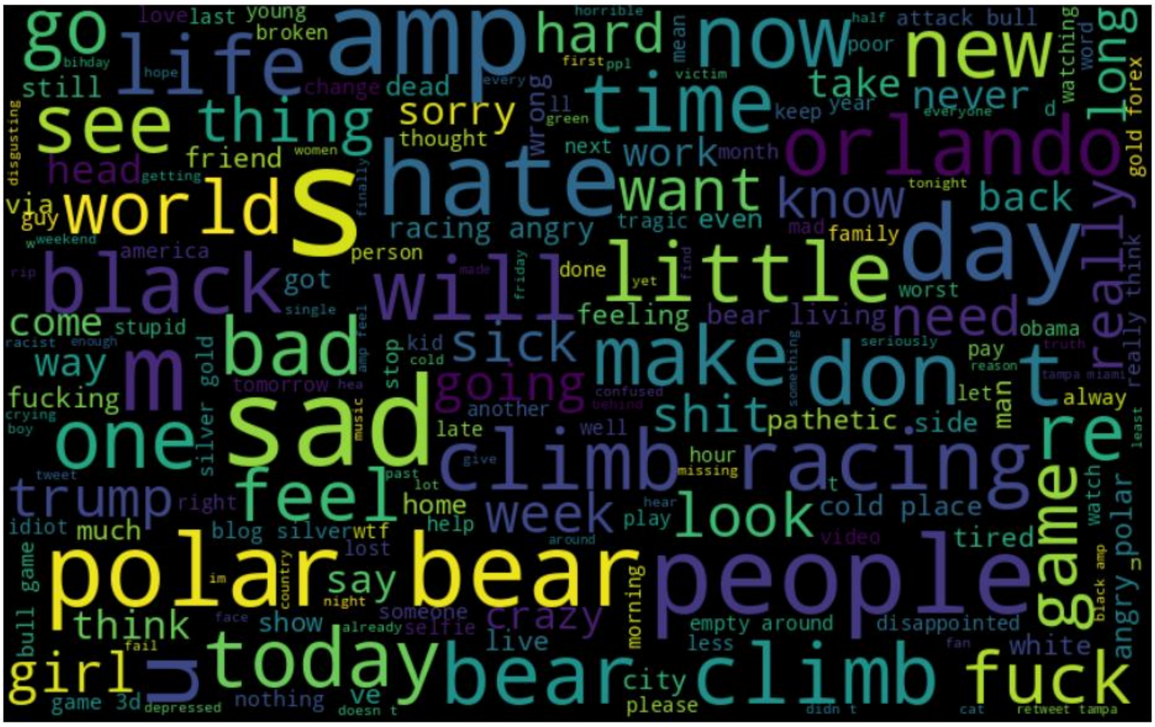

Figure 5. Negative Tweet

2.5 Stacking Ensemble Learning

The stacked ensemble classifier is one of the most popular classification models today. This approach uses the concept of a meta-learner that allows to find the best results by combining predictions from different base learning algorithms [16].The stacking ensemble learning technique in this research combines Support Vector Machine

(SVM), Decision Tree, and Random Forest models. Furthermore, the dataset is divided into training data and testing data. Before model building, at this stage a CountVectorizer is performed to convert text into numerical representations so that it can be used in machine learning algorithms. At this stage, relevant features are extracted from the tweet text in the data using methods such as TF-IDF (Term Frequency-Inverse Document Frequency) or word frequency counting using the sklearn library which provides the functions and algorithms of these methods.

The stacking technique starts with the creation of Support Vector Machine (SVM), Decision Tree, and Random Forest models. The models are then trained and tested with the split dataset and then evaluated using confusion matrix. Next, a new Meta-Learner combination model is created that combines the previously tested models. The Stacking Ensemble Learning technique is then trained and tested with the same dataset and evaluated using confusion matrix.

2.6 Evaluation

The evaluation stage is a stage that must be carried out in this research to measure the performance of Support Vector Machine, Decision Tree, and Random Forest in sentiment analysis of hate speech on Twitter. After the model is trained using training data, the evaluation stage is carried out using testing data that has previously been separated. The testing data is used to test the model's ability to predict hate speech sentiment by comparing the predicted results with the actual labels. Commonly used evaluation metrics include accuracy, recall, and F1-score [19]. The following is a brief explanation of the various evaluation metrics.

1. Accuracy: is the degree of precision of a measurement in relation to its true value. In mathematical notation, accuracy can be represented by Equation (1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Precision: is the ability of our model to predict good hate. Mathematically, precision can be represented by Equation (2).

$$Precision = \frac{TP}{TP + FP}$$

3. Recall: is the true positive rate, i.e. the ability of our model to detect all hate speech. Mathematically, it can be represented by Equation (3).

$$Recall = \frac{TP}{TP + FN}$$

4. F1-Score: evaluates the trade-off between recall and precision (harmonic mean). Mathematically, it can be represented by Equation (4).

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

In the evaluation matrix illustration above in the context of sentiment analysis of hate speech on twitter, True Positive (TP) represents a sample of tweets classified as hate speech, False Positive (FP) represents a sample of hate speech classified as hate speech, True Negative (TN) represents samples of hate speech that are classified as not hate speech, while False Negative (FN) represents samples of hate speech that are classified as not hate speech.

By comparing the model's prediction with the actual label, it can be evaluated to what extent the model can correctly classify hate speech. In addition, confusion matrix analysis is also conducted to see the extent to which the model can distinguish between hate speech and non-hate speech classes. With careful evaluation stages, this research can evaluate the performance and effectiveness of logistic regression and multinomial naive bayes approaches in sentiment analysis of hate speech on Twitter based on the Tweeter Hate Speech Analysis dataset.

RESULTS AND DISCUSSION

3.1 Compare Purposed Method With Base Method

The results of sentiment analysis using the SVM model get a fairly good accuracy value, which is 93.59% and F1-Score 93.48%. The results of the SVM model confusion matrix evaluation are shown in Figure 6.

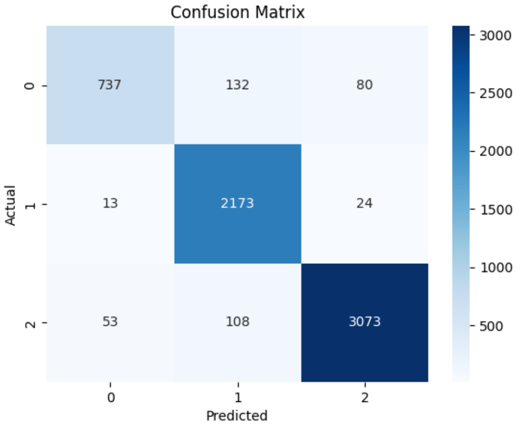

Figure 6. Confusion Matrix SVM

The results of sentiment analysis using the Decision Tree model get a fairly good accuracy value, which is 93.91% and F1-Score 93.83%. The results of the confusion matrix evaluation of the Decision Tree model are shown in Figure 7.

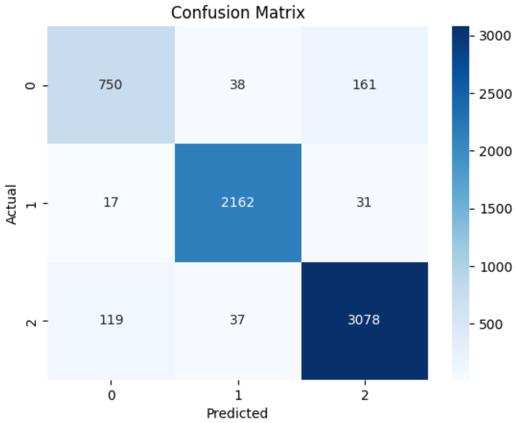

Figure 7. Confusion Matrix Decision Tree

The results of sentiment analysis using the Random Forest model get a fairly good accuracy value, which is 93.91% and F1-Score 93.83%. The results of the Random Forest model confusion matrix evaluation are shown in Figure 8.

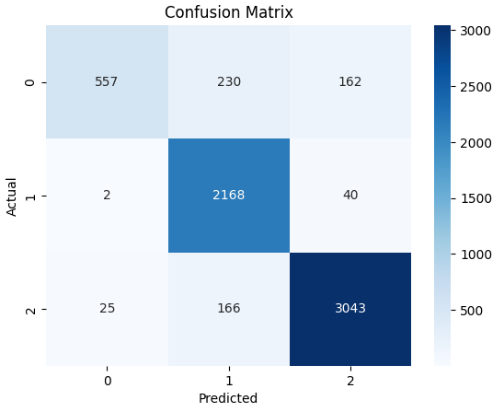

Figure 8. Confusion Matrix Random Forest

The Stacking Ensemble Learning technique obtained excellent results, namely with an accuracy value of 96.16% with an F1-Score of 96.13%. The confusion matrix evaluation results of the Stacking Ensemble Learning technique are shown in Figure 9.

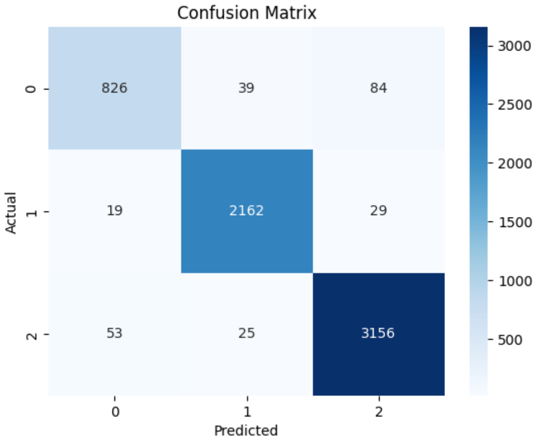

Figure 9. Confusion Matrix Stacking Ensemble Learning

The Stacking Ensemble Learning technique obtained excellent results with an accuracy value of 96.16% with an F1-Score of 96.13%. In this research, the Stacking Ensemble Learning technique is compared with basic and contemporary machine learning models. As a result, the proposed model, namely the stacking ensemble learning technique outperforms the previous models as shown in table 5.

Table 5. Comparison of Ensemble Learning Stacking Technique with Basic Model and Existing Models

Algorithms	Accuracy	Precision	Recall	F1 Score
SVM	93,58%	93,69%	93,59%	93,48%
Decision Tree	93,70%	93,59%	93,70%	93,63%
Random Forest	90,22%	90,82%	90,00%	89,70%
CNN [9]	79,00%	-	-	-
BERT[10]	75,00%	-	-	-
Random forest dan Autoencoder[4]	94,00%	-	-	-
Proposed Method Stacking Ensemble Learning (SVM, Decision Tree, Random Forest)	96,11%	96,07%	96,10%	96,07%

CONCLUSION

The results show that the Stacking Ensemble Learning technique provides better results compared to the basic model and other existing models. The SVM model gives an accuracy of 93.58%, Decision Tree 93.70%, and Random Forest gives an accuracy of about 90.22%. However, the stacking ensemble learning technique outperforms all these models with an accuracy of 96.16% and F1-Score of 96.13%. This shows that the use of Stacking Ensemble Learning technique can improve the accuracy of sentiment analysis of hate speech on Twitter.

Overall, this research shows that the Stacking Ensemble Learning technique has the potential to improve the accuracy of hate speech sentiment analysis on Twitter. This can contribute to creating a more positive and safe online environment with better ability to identify and address hate speech.

REFERENCES

[1] K. Olteanu, A., Castillo, C., Boy, J., & Varshneya, “Social media and hate speech: A machine learning perspective,” *Proc. 12th ACM Conf. Web Sci.*, pp. 279–287, 2019, doi: 10.1145/3292522.3326017.

[2] F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga, “Machine learning techniques for hate speech classification of twitter data: State-of-The-Art, future challenges and research directions,” *Comput. Sci. Rev.*, vol. 38, p. 100311, 2020, doi: 10.1016/j.cosrev.2020.100311.

[3] H. Madhu, S. Satapara, S. Modha, T. Mandl, and P. Majumder, “Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments,” *Expert Syst. Appl.*, vol. 215, no. May 2022, p. 119342, 2023, doi: 10.1016/j.eswa.2022.119342.

[4] I. R. Ammar Aouchiche, F. Boumahdi, A. Madani, and M. A. Remmide, “Hate Speech Prediction on Social Media,” *SN Comput. Sci.*, vol. 4, no. 3, pp. 1–7, 2023, doi: 10.1007/s42979-023-01668-6.

[5] C. Min *et al.*, “Finding hate speech with auxiliary emotion detection from self-training multi-label learning perspective,” *Inf. Fusion*, vol. 96, no. January, pp. 214–223, 2023, doi: 10.1016/j.inffus.2023.03.015.

[6] N. Tri Romadloni, I. Santoso, and S. Budilaksono, “Perbandingan Metode Naive Bayes, Knn Dan Decision Tree Terhadap Analisis Sentimen Transportasi Krl Commuter Line,” *J. IKRA-ITH Inform.*, vol. 3, no. 2, pp. 1–9, 2019.

[7] W. Hidayatullah, M. R., & Maharani, “Depression Detection on Twitter Social Media Using Decision Tree,” *J. RESTI*, vol. 6(4), pp. 677–683, 2022, [Online]. Available: http://jurnal.iaii.or.id

[8] D. A. Pisner, “Support Vector Machine,” *Mach. Learn.*, pp. 101–121, 2020, doi: 10.1016/B978-0-12-815739-8.00006-7.

[9] M. Siino, E. Di Nuovo, I. Tinnirello, and M. la Cascia, “Detection of Hate Speech Spreaders using convolutional neural networks,” *CEUR Workshop Proc.*, vol. 2936, pp. 2126–2136, 2021.

[10] E. Finogeev, M. Kaprielova, A. Chashchin, K. Grashchenkov, G. Gorbachev, and O. Bakhteev, “Hate speech spreader detection using contextualized word embeddings,” *CEUR Workshop Proc.*, vol. 2936, pp. 1937–1944, 2021.

[11] Muslim, M. A., Nikmah, T. L., Pertiwi, D. A. A., Subhan, Jumanto, Dasril, Y., and Iswanto., “New model combination meta-learner to improve accuracy prediction P2P lending with stacking ensemble learning: Intelligent Systems with Applications,” vol. 18, p. 200204, 2023, doi: 10.1016/j.iswa.2023.200204

[12] Sidik, D. D., & Sen, T. W. “Penggunaan Stacking Classifier Untuk Prediksi Curah Hujan,” IT FOR SOCIETY, vol. 04(01), 2021.

[13] Putri, A. K., & Suparwito, H. “Uji Algoritma Stacking Ensemble Classifier pada Kemampuan Adaptasi Mahasiswa Baru dalam Pembelajaran Online,” KONSTELASI: Konvergensi Teknologi dan Sistem Informasi, vol. 3(1), 1, 2023.

[14] Chen, E. L. E., Chen, C. Y. L., and Lee, C. Y. L. “Meta-rPPG: Remote Heart Rate Estimation Using a Transductive Meta-Learner,” arXiv preprint arXiv:2007.06786v1, 2020.

[15] AL-Alimi, D., Al-qaness, M. A. A., Cai, Z., Dahou, A., Shao, Y., & Issaka, S. “Meta-Learner Hybrid Models to Classify Hyperspectral Images. Remote Sensing,” vol. 14(4), 1038, 2022, doi: 10.3390/rs14041038.

[16] M. Ragab, A. M. Abdel Aal, A. O. Jifri, and N. F. Omran, "Enhancement of predicting students performance model using ensemble approaches and educational data mining techniques," Wireless Communications and Mobile Computing, vol. 2021, pp. 1-9, 2021.

[17] A. Ollagnier, E. Cabrio, and S. Villata, “Unsupervised fine-grained hate speech target community detection and characterisation on social media,” Soc. Netw. Anal. Min., vol. 13, no. 1, 2023, doi: 10.1007/s13278-023-01061-4.

[18] P. Kathiravan, R. Saranya, and S. Sekar, “Sentiment Analysis of COVID-19 Tweets Using TextBlob and Machine Learning Classifiers BT - Proceedings of International Conference on Data Science and Applications,” 2023, pp. 89–106.

[19] Ruuska, S., Hamäläinen, W., Kajava, S., Mughal, M., Matilainen, P., & Mononen, J., “Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle,” Behavioural Processes. 2018, doi: 10.1016/j.beproc.2018.01.004.