International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Journal Européen des Systèmes Automatisés
Vol. 58, No. 1, January, 2025, pp. 13-20
Journal homepage: http://iieta.org/journals/jesa

Optimizing Hate Speech Detection in Indonesian Social Media: An ADASYN and LSTM-Based Approach

Febby Apri Wenando[1,2], Nooraini Yusoff[2], Nurul Izrin[3], Sulistiawati R. N. Ahmad[4], M. Salim[4], Misrawati A. Puspa[5], Dony Novaliendry[6*]

[1] Department of Information Technology, Universitas Andalas, Padang 25163, Indonesia
[2] Department of Data Science, University Malaysia Kelantan, Kelantan 16100, Malaysia
[3] Department of Software Engineering, University Teknikal Malaysia, Melaka 76100, Malaysia
[4] Department of Information System, ITSBM Selayar, Sulsel 92812, Indonesia
[5] Department of Information System, Universitas Ichsan Gorontalo, Gorontalo 96138, Indonesia
[6] Department of Electronic Engineering, Universitas Negeri Padang, Padang 25131, Indonesia

Corresponding Author Email: dony.novaliendry@ft.unp.ac.id

https://doi.org/10.18280/jesa.580102

Received: 2 October 2024
Revised: 17 December 2024
Accepted: 17 January 2025
Available online: 31 January 2025

Keywords:
text mining, imbalanced class, oversampling, Adaptive Synthetic (ADASYN), LSTM

ABSTRACT

Identifying hate speech in Indonesian social media presents considerable difficulties owing to the intricacies of the language and the varied nature of online material. This paper presents a novel method for improving hate speech identification in Indonesia by tackling the significant class imbalance in Indonesian hate speech datasets. The ADASYN oversampling technique proficiently addresses this problem, representing a notable advancement in this study. The FastText method is utilized for word weighting, improving the prediction efficacy of the classification model. The dataset is carefully curated to authentically reflect the language characteristics and cultural circumstances of Indonesian social media conversation. The long short-term memory (LSTM) method is chosen for its capacity to record long-range relationships in sequential data, essential for comprehending the context of hate speech. The assessment of performance using criteria like accuracy, precision, recall, and F1-Score illustrates the efficacy of this method in precisely detecting hate speech. This research markedly enhances hate speech identification technology in Indonesian language processing, offering a viable method to curtail the dissemination of harmful information on internet platforms. The results of this study include practical implications for formulating more effective tactics to combat hate speech on Indonesian social media.

1. INTRODUCTION

Social media serves as a platform for individuals to communicate and share information. Popular social media platforms in Indonesia include Facebook, Instagram, and Twitter. While social media aims to facilitate human communication and the sharing of positive content, it also presents challenges. Every societal change brings both positive and negative aspects. While these platforms provide freedom of expression, they can also be misused to spread hate speech [1].

Hate speech refers to any act that denigrates or attacks individuals or groups. It can manifest as provocation, slander, or insults, often targeting race, skin color, gender, or religion, among other aspects [2]. Indonesia lacks an automated system for detecting hate speech, making it challenging to address such content on social media. However, Indonesia has laws, such as Law Number 11 of 2008 on Electronic Information and Transactions (UU-ITE), that regulate hate speech.

The growing prevalence of hate speech is particularly noticeable during national events, such as presidential elections, where supporters of opposing candidates engage in online hate speech and smear campaigns [3]. Currently, hate speech perpetrators can only be prosecuted through manual reporting to law enforcement. This manual process means perpetrators may go unpunished for verbal crimes without reports [4].

Hate speech evolves, adapting to new languages and vocabularies, making it difficult for automated systems to detect and filter such content. This necessitates adaptive detection and filtering systems to keep pace with these changes. Machine learning methods are often used in processing such data, where the collected data is used to make predictions and detect patterns over time.

However, one challenge in machine learning models is data's changing distribution and dispersion over time [5-7]. For example, the model must adapt to changing behavior patterns in studying customer behavior for sales predictions. Moreover, the class distribution in datasets can affect the performance of prediction models, particularly in unbalanced datasets.

In Indonesia, detecting hate speech requires an automatic system to identify such content effectively [8, 9]. However, the current Indonesian language hate speech dataset is limited, comprising only 13,169 samples from tweets on Twitter. This dataset [10] underwent a lengthy and challenging process. After collection, the dataset was suboptimal, with an unbalanced class distribution, resulting in weak performance for hate speech detection.

The lack of data, along with uneven class distribution, impedes the efficacy of machine learning algorithms in identifying hate speech [11, 12]. This study introduces the ADASYN method to equilibrate class distribution in the Indonesian language hate speech dataset to tackle this issue. The efficacy of this method is evaluated by the LSTM algorithm to enhance hate speech identification. This project seeks to enhance hate speech identification technology in Indonesian language processing, providing a viable remedy to reduce the dissemination of harmful information on internet platforms. Moreover, the results of this study possess practical significance for formulating more efficacious measures to combat hate speech on Indonesian social media.

2. LITERATURE REVIEW

This section will conduct a comprehensive literature study of scientific theories relevant to our research topic. We aim to gain a deeper understanding of key concepts, developments, and discoveries related to our research problem or field of study. This analysis will provide a detailed summary of existing theories and research conducted on our subject, as well as any areas that require additional research to address gaps in this research topic. Machine Learning is a part of artificial intelligence that allows systems to learn automatically and improve abilities based on experience without reprogramming. This system can learn independently through data and learn from that data, one of which is an automatic approach that is used to detect hate speech.

Numerous researchers are actively advancing the development of hate speech detection methodologies [13-20]. In the context of hate speech in Indonesian, the situation has not yet achieved optimal levels, primarily due to the limited availability of Indonesian language datasets [2, 9, 10, 18, 21, 22]. The most recent dataset for the Indonesian language was compiled by the study [10]. The research encompasses a total of 13,169 datasets in text form derived from tweets on Twitter. It has undergone multiple phases to ensure the dataset's validity, beginning with data collection through a crawling process. This was followed by Focus Group Discussions (FGD) involving personnel from the Direktorat Tindak Pidana Siber Badan Reserse Kriminal Kepolisian Negara Republik Indonesia (BARESKIM POLRI), consultations with language experts, and the implementation of crowdsourcing for the annotation process [23, 24]. The dataset from this research requires a more balanced distribution, with 7000 instances of positive classes compared to 5000 instances of negative classes. The imbalance present in this dataset is likely to influence the performance during data processing, resulting in an accuracy value of 77.36%. The current outcomes remain suboptimal. The collected datasets require balance across each class, leading to suboptimal performance of all machine learning algorithms in developing models for automatic hate speech detection systems. Consequently, there is a necessity for a classification model capable of optimizing the distribution within an unbalanced dataset. The research was carried out to balance the dataset using resampling techniques [6, 25-30]. The Oversampling approach changes sample data by adding sample data contained in the minority class by making replicas of the sample data until the distribution of sample data becomes more balanced. On the other hand, undersampling changes the sample data by eliminating sample data in the majority class until the sample data distribution becomes more balanced. However, oversampling techniques are more widely used for tasks with imbalanced dataset classes because oversampling maximizes existing datasets by creating new synthetic classes.

The Synthetic Minority Over-sampling Technique (SMOTE) is a widely utilized method for mitigating class imbalance by oversampling. Nonetheless, SMOTE possesses a significant limitation: it produces synthetic observations without accounting for the unique attributes of the minority class examples. As a result, this may cause substantial changes in the class borders between majority and minority classes, thereby distorting the original data distribution and inadequately representing the intrinsic features of the minority class. He et al. [31] presented the Adaptive Synthetic (ADASYN) to resolve this issue. ADASYN is intended to provide synthetic situations that reflect the learning difficulties linked to specific minority class observations. It produces additional synthetic examples for minority class observations that are particularly difficult to learn, so accommodating the unique qualities and challenges of each instance in the minority class. This method seeks to deliver a more equitable and precise depiction of the minority class while maintaining the original data's distribution.

The comparative literature between the two has been widely carried out [25, 32-34], with the advantage that ADASYN produces synthetic observations along a straight line between minority class observations and their k nearest minority class neighbors. Thus, ADASYN was designed to handle some of the problems of SMOTE. The pseudocode of ADASYN technique is shown in Table 1.

Table 1. Pseucode of ADASYN technique [25]

```
Algorithm 1: ADASYN
Input:
  Training dataset $X_r$, Hyper parameter $\beta \in [0,1]$, $K$=5
  The $i$ th sample in the minority class $x_i$ ($i$ = 1,2,3, ... , $m_s$),
  A random minority sample $x_{zi}$ in K-nearest neighbors of $x_i$.
Output:
  Synthetic minority samples $s_i$, Oversampled training dataset $X_{ADASYN}$

Calculate the number of majority samples $m_l$ and the number of minority samples $m_s$ in the training dataset $X_T$
According to the formula $G = (m_i - m_s) \times \beta$, calculates the number of samples to be synthesized for the minority class.
For each example $xi \in$ minority class:
    Calculate $\Delta i$  //the number of majority samples in K-nearest neighbors of minority Sample $x_i$
    Calculate $r_i = \Delta i/K$  //the ratio of majority samples in K-nearest neighbors of minority Sample $x_i$
    Standardize $r_i$ through the formula $\widehat{r_l} = r_i / \sum_{i=1}^{m_s} r_i$
    Calculate $g_i = \widehat{r_l}$ x $G$  //the number of new samples to be generated for each minority $x_i$
    Do the lopp from 1 to $g_i$
        Using the formula $s_i = x_i + (x_{zi} - x_i)$ x $y$ to synthesize data samples  //$y$ is a random number: $y \in |$
    End
End
Return Oversamples training datasets $X_{ADASYN}$
```

The proposed ADASYN and LSTM-based approach for hate speech detection in Indonesian social media distinguishes itself from prior methods in several significant aspects. Prior research encountered difficulties due to imbalanced datasets, adversely affecting the performance of machine learning models. Traditional oversampling techniques, such as SMOTE (Synthetic Minority Over-sampling Technique), face criticism for distorting class boundaries due to the generation of synthetic samples that inadequately represent the characteristics of the minority class [35, 36]. The ADASYN technique enhances this process by generating synthetic samples that focus on challenging instances within the minority class, thereby achieving a more refined balance in the dataset. This study advances prior research by incorporating TF-IDF weighting to improve feature representation, beyond standard preprocessing techniques [37]. This guarantees that the textual data input to the LSTM model is organized and significant, which is essential for efficient sequential data learning. The use of LSTM models, in contrast to traditional machine learning models employed in previous research, enables the effective capture of long-range dependencies and contextual relationships in text data, which are crucial for comprehending nuanced hate speech [38]. This is especially advantageous considering the linguistic intricacies of Indonesian social media discourse.

By leveraging ADASYN and LSTM together, this study not only mitigates the common pitfalls of dataset imbalance and insufficient contextual modeling but also sets a new benchmark for hate speech detection in Indonesian language processing. These improvements highlight the practicality of this method for real-world applications.

3. RESEARCH METHODS

This study carried out several stages in the Adaptive Synthetic (ADASYN) oversampling process to overcome imbalanced data on hate speech detection as shown in Figure 1.

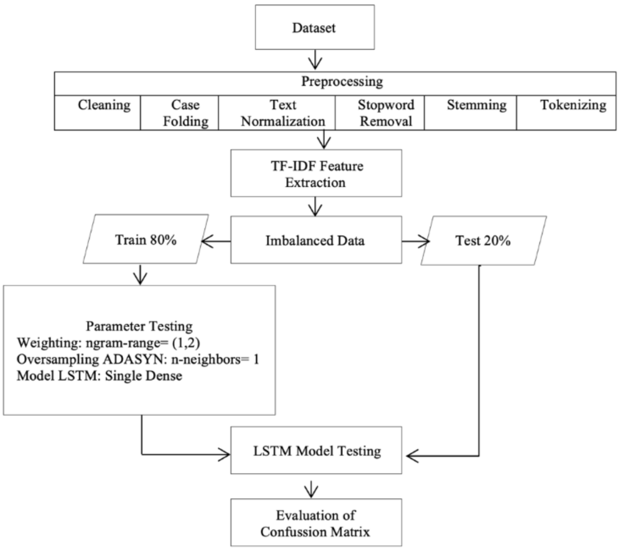


Figure 1. Flowchart

The study uses a dataset for the detection of hate speech in Indonesian that has been provided by the study [10]. Ibrohim and Budi [10] collected 13169 tweets in Indonesian. In the dataset, there are 2 columns, namely Label and Twit. In Label, tweet data is grouped into 2, namely Non_HS and HS. Non_HS for non-hate speech and HS for hate speech. The data that has been successfully collected is preprocessed to remove unnecessary noise in the classification process. After the dataset has been cleaned, the next step is word weighting with the TF-IDF feature.

The Oversampling Adaptive Synthetic (ADASYN) approach was utilized in this study, and the algorithm that was utilized was a machine learning algorithm such as Long Sort Term-Memory (LSTM). Both of these processes were utilized in this research. The Confusion Matric was utilized in order to classify the LSTM Model that was applied to the dataset. Confusion Matrix is a test method that is used to calculate the level of accuracy and compare it with the results of the classification of the model that is used in the form of data tables. After then, the outcomes of the categorization will be compared with the data that was really collected. By producing synthetic samples for the minority class (Hate Speech) depending on the difficulty of learning specific instances, the ADASYN (Adaptive Synthetic Sampling) approach is utilized in order to correct the imbalance that exists within the dataset including hate speech. This method makes use of the n_neighborsn_neighbors parameter in order to determine the closest neighbors for every instance of the minority class. Furthermore, it gives larger weights to instances that have a greater number of neighbors belonging to the majority class. This method focuses the production of synthetic data on places that are more difficult to learn.

After that, an LSTM model was applied to the oversampled data in order to detect hate speech, leading to improved performance measures such as F1-Score, recall, accuracy, and precision. Various models were tested to determine the effect of tuning the method’s hyperparameters, such as n_neighbors and random_state. Results showed that n_neighbors=1 and n_neighbors=2 both improved the model's output. The adaptive oversampling played a crucial role in improving the LSTM classifier's prediction accuracy and reducing the impact of unbalanced datasets.

4. RESULTS AND DISCUSSIONS

The author utilizes a dataset sourced from the study [10]. The dataset comprises 13 columns, including 1 Tweet column and 12 multi-column labels. This research utilizes a dataset comprising two columns: Tweet and HS. The Tweet column is a string data type that comprises sentences containing letters, numbers, and symbols. The HS column comprises two labels, 1 and 0, represented as integer data types. This dataset comprises 5,561 labels for hate speech tweets and 7,608 labels for non-hate speech tweets. The HS column revealed a label imbalance, specifically indicating that the quantity of hate speech data is less than that of non-hate speech data, with a discrepancy of 2047. Additionally, data analysis involved data exploration, which included the visual description, explanation, and extraction of information from the dataset. This process aims to analyze the structure of the data for processing.

The ADASYN and LSTM-based approach excels in detecting Indonesian hate speech due to its ability to address key challenges in the language and social media context. Indonesian social media often features informal language, slang, and mixed dialects, making hate speech detection complex. The LSTM algorithm effectively captures long-range dependencies and contextual nuances in sequential data, enabling the model to understand subtle expressions of hate speech. Additionally, the ADASYN oversampling method handles class imbalance in the dataset by generating synthetic samples for underrepresented hate speech cases, ensuring better performance and reducing bias in classification.

Furthermore, Indonesian hate speech frequently revolves around culturally sensitive topics like race, religion, and politics, and the curated dataset reflects these nuances. Preprocessing steps such as cleaning, normalization, and TF-IDF weighting enhance the quality of inputs, while LSTM's adaptability to evolving language patterns on social media ensures robustness against changes in vocabulary and expression styles. This combination makes the approach particularly effective in addressing the unique linguistic and contextual challenges of hate speech detection in Indonesia.

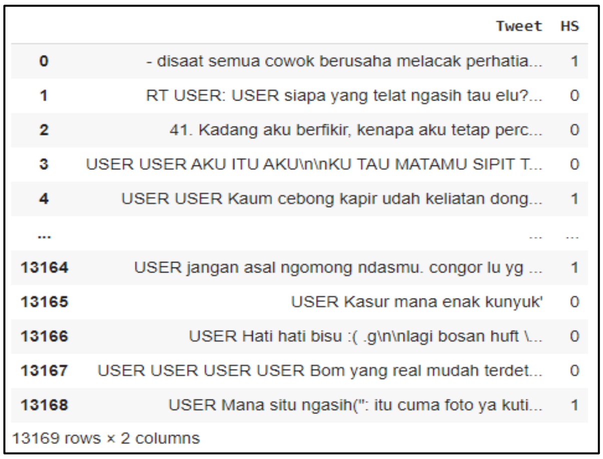
4.1 Pre-processing

This stage is carried out to prepare the dataset before it is used to train the classification model. Pre-processing is carried out in six sub-processes that have their respective functions. The preprocessing stage is carried out to prepare the dataset before it is used to train the classification model. Pre-processing is carried out in six sub-processes that have their respective functions. The preprocessing results, as shown in Table 2, illustrate the transformation of raw text into processed data suitable for classification.

Figure 2 illustrates the changes in the dataset before and after preprocessing, emphasizing the removal of noise and normalization of text data.

Table 2. Preprocessing results

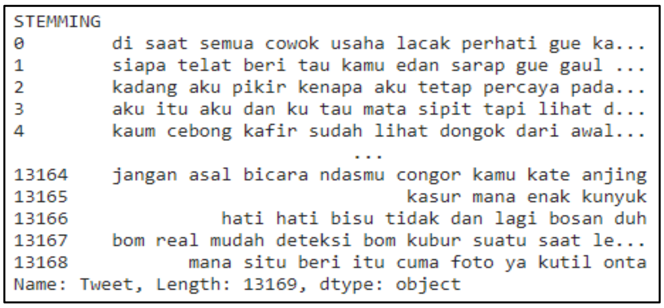
Preprocessing	Input	Output
Cleaning	RT USER USER siapa yang telat ngasih tau elu?	RT USER USER siapa yang telat ngasih tau elu
Case Folding	RT USER USER siapa yang telat ngasih tau elu	rt user user siapa yang telat ngasih tau elu
Normalisasi	rt user user siapa yang telat ngasih tau elu	siapa yang telat memberi tau kamu
Stopword Removal	siapa yang telat memberi tau kamu	siapa telat memberi tau kamu edan
Stemming	siapa telat memberi tau kamu edan	siapa telat beri tau kamu
Tokenizing	siapa telat beri tau kamu	'siapa' 'telat' 'beri' 'tau' 'kamu'



	Tweet	HS
0	- disaat semua cowok berusaha melacak perhatia...	1
1	RT USER: USER siapa yang telat ngasih tau elu?...	0
2	41. Kadang aku berfikir, kenapa aku tetap perc...	0
3	USER USER AKU ITU AKU\n\nKU TAU MATAMU SIPIT T...	0
4	USER USER Kaum cebong kapir udah keliatan dong...	1
...	...	...
13164	USER jangan asal ngomong ndasmu. congor lu yg ...	1
13165	USER Kasur mana enak kunyuk'	0
13166	USER Hati hati bisu :(.g\n\nlagi bosan huft \...	0
13167	USER USER USER USER Bom yang real mudah terdet...	0
13168	USER Mana situ ngasih(": itu cuma foto ya kuti...	1

13169 rows × 2 columns

(a)



```
STEMMING
0        di saat semua cowok usaha lacak perhati gue ka...
1        siapa telat beri tau kamu edan sarap gue gaul ...
2        kadang aku pikir kenapa aku tetap percaya pada...
3        aku itu aku dan ku tau mata sipit tapi lihat d...
4        kaum cebong kafir sudah lihat dongok dari awal...
                                ...
13164    jangan asal bicara ndasmu congor kamu kate anjing
13165                                kasur mana enak kunyuk
13166              hati hati bisu tidak dan lagi bosan duh
13167    bom real mudah deteksi bom kubur suatu saat le...
13168           mana situ beri itu cuma foto ya kutil onta
Name: Tweet, Length: 13169, dtype: object
```

(b)

Figure 2. Pre-processed dataset: (a) Before (b) After

4.2 Weighting TF-IDF

Before entering the LSTM, Next is the TF-IDF weighting process code in the Tweet column data row. Datasets that have been preprocessed to preprocessing stemming will be weighted using the TF-IDF word weighting with the Eqs. (1)-(3).

TF-IDF score for the word t in the document d from the document set D is calculated as follows:

$$tf\ idf(t,d,D) = tf(t,d).idf(t,D) \tag{1}$$

$$tf(td) = \log(1 + freq(t,d)) \tag{2}$$

where,

$$idf(t,D) = \log\frac{N}{count(d \in D{:}\, t\ \in d)} \tag{3}$$

The results of this process, summarized in Table 3, highlight the calculated weights for terms in the dataset, which play a critical role in enhancing feature representation.

Table 3. TF-IDF matrix result

Index	Weight
0, 27608	0.18033391939380292
0, 61550	0.18033391939380292
0, 14926	0.18033391939380292
0, 61473	0.18033391939380292
0, 64559	0.16010251207748072
0, 67564	0.18033391939380292
0, 63221	0.18033391939380292
0, 43794	0.15069544338853197
0, 104902	0.18033391939380292

The TF-IDF matrix presented consists of rows that denote individual documents from the dataset, while the columns represent distinct terms within the complete corpus. The matrix values denote the TF-IDF weights allocated to each term within each document. TF-IDF (Term Frequency-Inverse Document Frequency) is a quantitative measure that indicates the significance of a term within a document in relation to a set of documents (corpus).

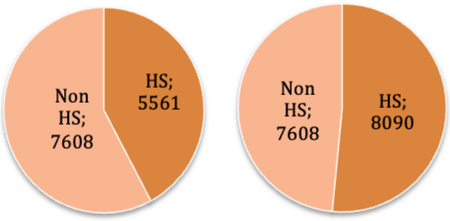
In the first row of the matrix, each entry represents the TF-IDF weight of a particular term in the first document. The entry (0, 27608) with a value of 0.18033391939380292 represents the TF-IDF weight of the term located at index 27608 in the first document. Higher TF-IDF weights signify greater importance of a term within a specific document and throughout the overall dataset.

TF-IDF computation identifies salient or distinctive terms in individual documents, facilitating text mining, information retrieval, and document classification.

4.3 Oversampling ADASYN imbalance dataset

Oversampling is done to balance the number of minority classes, namely positive hate speech, so that it is equal to negative hate speech, then a class balancing technique is used on the dataset, namely ADASYN (Adaptive Synthetic), Below is the output of the dataset that has been handled by imbalance.

According to the code provided, the number of negative hate speech data lines is now 7608, while the number of positive hate speech data lines is 8090. Before the ADASYN procedure, the numbers were 5561 and 7608, respectively. Figure 3 highlights the effect of this approach on the dataset by comparing the distribution of classes before and after balancing. Out of 13,169 tweets in the original dataset, 7,000 were deemed to be non-hate speech and 5,000 were deemed to be hate speech [10]. What this shows is that the dataset was already unbalanced before processing began.



(a) (b)

Figure 3. ADASYN processing: (a) Before (b) After

After cleaning and preprocessing the dataset for this research, the reported count is 5,561 hate speech tweets and 7,608 non-hate speech tweets. This suggests that some data points were either filtered out or transformed during preprocessing, resulting in adjusted totals. Preprocessing likely included steps such as removing noise, irrelevant entries, or duplicates, which can explain the difference.

The difference arises due to the cleaning and preprocessing procedures, which are crucial to preparing the dataset for machine learning tasks. The preprocessing stage refined the dataset to ensure quality and relevance, leading to slightly different numbers than the original dataset.

4.4 LSTM

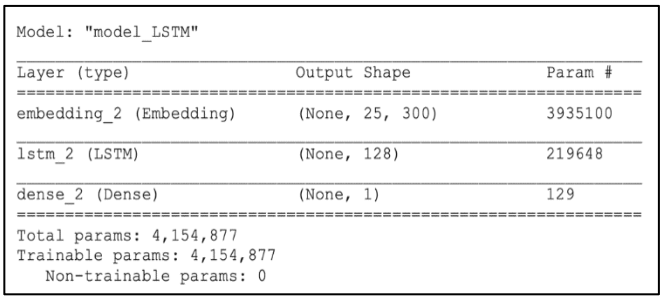
The LSTM model is constructed via a layered architecture through a Sequential function, including one input tensor and one output tensor. The input tensor denotes the matrix obtained from the prior TF-IDF weighting outcomes. The following code constructs the LSTM classification model. Upon the creation of the classification model, it is assembled for use in the model training phase. The modified parameters are as follows:

1) The loss function used is ‘binary cross-entropy’ since the predicted data is in binary form.

2) The optimizer used is the ADAM Optimizer.

3) Metrics include ‘accuracy’.

Compiling the model displays the model’s dimensions, as shown in Figure 4.



```
Model: "model_LSTM"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_2 (Embedding)      (None, 25, 300)           3935100
_________________________________________________________________
lstm_2 (LSTM)                (None, 128)               219648
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 129
=================================================================
Total params: 4,154,877
Trainable params: 4,154,877
   Non-trainable params: 0
```

Figure 4. LSTM model compile results

After the model is compiled, then fit the model with the dataset to train the model. The dataset that has been weighted and oversampled is first converted to a numpy array, as well as inputting some data to be used as test data. The adjusted model training parameters include:

1) The number of epochs is the number of trainings carried out as many as 100 epochs. During training, the accuracy trend over 100 epochs is visualized in Figure 5, indicating consistent improvements as the model converges.

2) The number of batch sizes is the number of data sets trained in the training as much as 250. Similarly, the reduction in loss values across training epochs, as shown in Figure 6, demonstrates the model’s progression towards optimal performance.

3) Multiprocessing to train the model in thread mode is enabled.

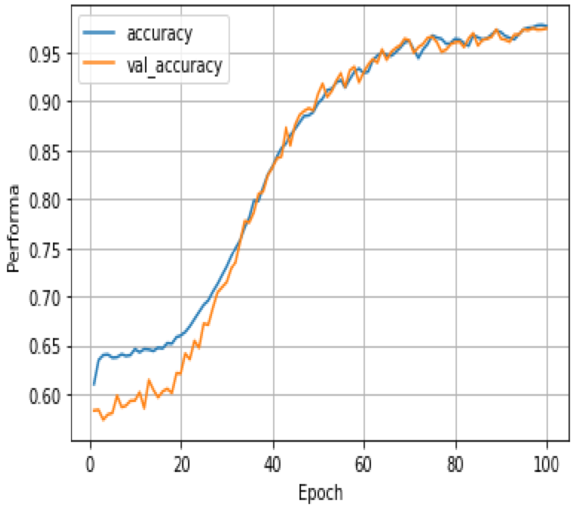


Figure 5. Chart accuracy value

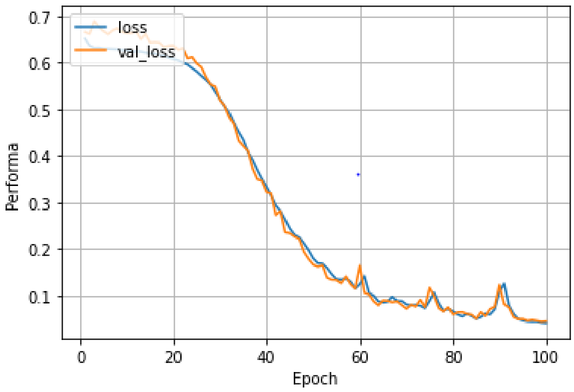


Figure 6. Chart loss value

Next is shown in the graph the accuracy value and loss value for each epoch on each training data and test data on the results of testing at 100 epochs.

The assessment phase is the concluding stage for assessing the model's performance and obtaining performance outcomes. The assessment outcomes present the model's performance as a confusion matrix table. The outcomes of this phase, illustrated in Table 4, indicate the efficacy of the suggested model for precision, recall, F1-score, and accuracy. For evaluation, the model is provided with data for prediction, and the projected outcomes are juxtaposed with the actual values. The model is provided with data that has been analyzed up to 20% of the overall training dataset.

Table 4. Result perform

Result				
	Precision	**Recall**	**F1-Score**	**Support**
0	0.94	0.92	0.93	1805
1	0.91	0.93	0.92	1334
accuracy			0.92	3139
macro avg	0.92	0.93	0.92	3139
weighted avg	0.92	0.92	0.92	3139
FINISH				

At 128 units per layer, the LSTM determines the capacity of the model and the size of the hidden state vector. For this purpose, we employ a sigmoid activation function and set the dropout rate to 0.2 to avoid overfitting.

The following are the outcomes of the LSTM model's assessment.

The following values are displayed in the confusion matrix image: 1.312 for true positive, 1.747 for false positive, 58 for false negative, and 22 for true negative. Accuracy, precision, recall, and F1-Score may then be determined from these numbers, and the result is a score: How near the projected value is to the actual value (the real value) is what we mean when we talk about accuracy.

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

$$Precission\ \% = \frac{TP}{TP + FP}$$

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

$$F1 - Score\ \% = 2 * \frac{Presisi * Recall}{Presisi + Recall}$$

4.5 Evaluation result

To get a conclusion from the overall model, the classification and evaluation process is carried out four times with several different parameters for each model, while the differences in model parameters are as follows:

1) TF-IDF weighting uses the parameter ngram_range = (1, 2) except in the 3rd model using the default parameters.
2) ADASYN uses the parameter n_neighbors =2 in models 1 and 2, while in model 4 it uses n_neighbors =1. Random_state in model 1 is 2, in models 2 and 4 is 0. While Model 3 uses the default ADASYN parameter.
3) Models 1 and 2 use 2 dense layers while models 3 and 4 use 1 dense.

So that it can be seen how big the difference in the results of the model when it is trained and retested on the 4 models that have been made. The results of the evaluation for the 4 tested models are as shown in Tables 5, 6, and 7.

Table 5. Training results

Model No.	Training Results			
	Loss	**Accuracy**	**Val Loss**	**Val Accuracy**
1	0,1	91 %	0,1	91%
2	0,1	90%	0,1	90%
3	0,0	90%	0,0	89%
4	0,0	92%	0,0	92%

Table 6. Predictions results

Model No.	Predictions Results			
	TP	**FP**	**TN**	**FN**
1	1203	66	1599	36
2	1190	80	1585	49
3	-	-	-	-
4	1312	58	1747	22

Table 7. Evaluations

Model No.	Evaluations						
	Accuracy	Precision		Recall		F1-Score	
		0	1	0	1	0	1
1	91%	93%	90%	91%	92%	92%	91%
2	90%	92%	89%	90%	91%	91%	90%
3	89%	-	-	-	-	-	-
4	92%	94%	91%	92%	93%	93%	92%

The study’s findings highlight practical implications for addressing hate speech on Indonesian social media by leveraging insights into linguistic patterns and contextual nuances. For social media platforms, this means developing advanced hate speech detection models tailored to the Indonesian language and cultural context, incorporating machine learning techniques that can differentiate hate speech from satire or benign content. Additionally, platforms can implement real-time moderation systems using these enhanced models to identify and address hate speech more effectively. Policymakers can use these insights to draft regulations requiring platforms to adopt culturally sensitive algorithms and promote digital literacy programs, encouraging users to recognize and counter hate speech responsibly. These combined efforts can help create safer and more inclusive online environments.

5. CONCLUSION

This study examined the challenge of detecting hate speech in Indonesian social media by proposing a new method that addresses the significant class imbalance present in current datasets. The integration of the Adaptive Synthetic (ADASYN) oversampling technique with the Long Short-Term Memory (LSTM) model resulted in notable enhancements in classification accuracy and reliability for hate speech detection. The implementation of ADASYN oversampling contributed to dataset balance, facilitating improved classification of both hate speech and non-hate speech instances by the model. The LSTM model, recognized for its capacity to capture long-range dependencies in sequential data, was essential in analyzing the contextual nuances of hate speech.

Based on the study's results, which involved classifying sentiment analysis on Twitter using a dataset of 13,169 entries with an imbalanced distribution between negative and positive hate speech, the following conclusion can be drawn: (1). At epoch 100, the LSTM learning machine model achieved an accuracy of 0.93% and a validation accuracy of 0.92%. 2. The classification results for hate speech demonstrate strong performance when employing the ADASYN oversampling technique in conjunction with the LSTM model. The accuracy achieved is 92%, with a negative precision of 94%, negative recall of 92%, and a negative F1-Score of 93% for hate speech. Additionally, the positive precision stands at 91%, positive recall at 93%, and the positive F1-Score at 92%.

This study provides significant contributions to the domain of hate speech detection. It offers a solid framework for addressing class imbalance in text classification tasks, particularly within Indonesian social media discourse. The ADASYN technique employed in this study addresses the limitations of conventional oversampling methods such as SMOTE, maintaining a realistic data distribution while improving the minority class for enhanced learning efficacy. The study's emphasis on Indonesian-specific datasets, linguistic features, and cultural context provides important insights for the development of automated systems that are adapted to regional and linguistic nuances. This research presents practical implications for the development of automated hate speech moderation tools, which may be implemented on social media platforms to identify and address harmful content in real time.

Future research may explore several potential avenues. The expansion of the dataset to incorporate a wider and more diverse array of social media data is crucial, as it would improve the model's generalizability and applicability to various online platforms. The advancement of real-time hate speech detection systems may further illustrate the model's efficacy in dynamic online contexts. Future research may investigate the extension of the framework to accommodate multilingual datasets or the integration of multimodal analysis, which includes text, images, and audio, to enhance the detection system's comprehensiveness. Additionally, there exists an opportunity to enhance the model to more effectively consider context and sentiment, facilitating more nuanced interpretations of hate speech. Ethical considerations must be prioritized, especially in reducing biases in dataset annotation and ensuring that the model does not unintentionally reinforce stereotypes or unfairly target particular groups. This research can develop into a more versatile and effective tool in the ongoing effort to combat online hate speech by exploring these directions.

REFERENCES

[1] Ginting, P.S.B., Irawan, B., Setianingsih, C. (2019). Hate speech detection on Twitter using multinomial logistic regression classification method. In 2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS), Bali, Indonesia, pp. 105-111. https://doi.org/10.1109/IoTaIS47347.2019.8980379

[2] Alfina, I., Mulia, R., Fanany, M.I., Ekanata, Y. (2017). Hate speech detection in the Indonesian language: A dataset and preliminary study. In 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Bali, Indonesia, pp. 233-238. https://doi.org/10.1109/ICACSIS.2017.8355039

[3] Wenando, F.A., Hayami, R., Novermahakim, A.Y. (2020). Tweet sentiment analysis for 2019 Indonesia presidential election results using various classification algorithms. In 2020 1st International Conference on Information Technology, Advanced Mechanical and Electrical Engineering (ICITAMEE), Yogyakarta, Indonesia, pp. 279-282. https://doi.org/10.1109/ICITAMEE50454.2020.9398513

[4] Hakim, L., Kusumasari, T.F., Lubis, M. (2018). Text mining of UU-ITE implementation in Indonesia. Journal of Physics: Conference Series, 1007(1): 012038. https://doi.org/10.1088/1742-6596/1007/1/012038

[5] Lee, C.Y., Yang, M.R., Chang, L.Y., Lee, Z.J. (2010). A hybrid algorithm applied to classify unbalanced data. In the 6th International Conference on Networked Computing and Advanced Information Management, Seoul, Korea (South), pp. 618-621.

[6] Jafarigol, E., Trafalis, T. (2023). A review of machine learning techniques in imbalanced data and future trends. arXiv Preprint arXiv: 2310.07917. https://doi.org/10.48550/arXiv.2310.07917

[7] Johnson, J.M., Khoshgoftaar, T.M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6(1): 1-54. https://doi.org/10.1186/s40537-019-0192-5

[8] Sutejo, T.L., Lestari, D.P. (2018). Indonesia hate speech detection using deep learning. In 2018 International Conference on Asian Language Processing (IALP), Bandung, Indonesia, pp. 39-43. https://doi.org/10.1109/IALP.2018.8629154

[9] Wenando, F.A., Fuad, E. (2019). Detection of hate speech in Indonesian language on Twitter using machine learning algorithm. Prosiding CELSciTech, 4: 6-8.

[10] Ibrohim, M.O., Budi, I. (2019). Multi-label hate speech and abusive language detection in Indonesian Twitter. In Proceedings of the Third Workshop on Abusive Language Online, pp. 46-57. https://doi.org/10.18653/v1/W19-3506

[11] Padurariu, C., Breaban, M.E. (2019). Dealing with data imbalance in text classification. Procedia Computer Science, 159: 736-745. https://doi.org/10.1016/j.procs.2019.09.229

[12] Ayo, F.E., Folorunso, O., Ibharalu, F.T., Osinuga, I.A. (2020). Machine learning techniques for hate speech classification of Twitter data: State-of-the-art, future challenges and research directions. Computer Science Review, 38: 100311. https://doi.org/10.1016/j.cosrev.2020.100311

[13] Ali, M.Z., Rauf, S., Javed, K., Hussain, S. (2021). Improving hate speech detection of Urdu tweets using sentiment analysis. IEEE Access, 9: 84296-84305. https://doi.org/10.1109/ACCESS.2021.3087827

[14] Aljundi, I.I., Novaliendry, D., Hendriyani, Y., Syafrijon, S. (2024). Mobile-based skin cancer classification system using convolutional neural network. Data and Metadata, 3: 649.

[15] d'Sa, A.G., Illina, I., Fohr, D. (2020). Classification of hate speech using deep neural networks. Revue

d'Information Scientifique & Technique, 25(1). https://hal.science/hal-03101938v1.

[16] Jiang, L., Suzuki, Y. (2019). Detecting hate speech from tweets for sentiment analysis. In 2019 6th International Conference on Systems and Informatics (ICSAI), Shanghai, China, pp. 671-676. https://doi.org/10.1109/ICSAI48974.2019.9010578

[17] Kovács, G., Alonso, P., Saini, R. (2021). Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources. SN Computer Science, 2(2): 95. https://doi.org/10.1007/s42979-021-00457-3

[18] Mansur, Z., Omar, N., Tiun, S. (2023). Twitter hate speech detection: A systematic review of methods, taxonomy analysis, challenges, and opportunities. IEEE Access, 11: 16226-16249. https://doi.org/10.1109/ACCESS.2023.3239375

[19] Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. arXiv Preprint arXiv: 1701.08118. https://doi.org/10.17185/duepublico/42132

[20] Ridhani, D., Krismadinata, Novaliendry, D., Ambiyar, Effendi, H. (2024). Development of an intelligent learning evaluation system based on big data. Data and Metadata, 3: 569. https://doi.org/10.56294/dm2024.569

[21] Amalia, A., Sitompul, O.S., Nababan, E.B., Mantoro, T. (2020). An efficient text classification using fastText for Bahasa Indonesia documents classification. In 2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA), Medan, Indonesia, pp. 69-75. https://doi.org/10.1109/DATABIA50434.2020.9190447

[22] Koto, F., Rahimi, A., Lau, J.H., Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. arXiv Preprint arXiv: 2011.00677. https://doi.org/10.48550/arXiv.2011.00677

[23] Lubis, A.R., Nasution, M.K. (2023). Twitter data analysis and text normalization in collecting standard word. Journal of Applied Engineering and Technological Science (JAETS), 4(2): 855-863. https://doi.org/10.37385/jaets.v4i2.1991

[24] Sabou, M., Bontcheva, K., Derczynski, L., Scharl, A. (2014). Corpus annotation through crowdsourcing: Towards best practice guidelines. In LREC, pp. 859-866.

[25] Chen, Z., Zhou, L., Yu, W. (2021). ADASYN-random forest based intrusion detection model. In Proceedings of the 2021 4th International Conference on Signal Processing and Machine Learning, pp. 152-159. https://doi.org/10.1145/3483207.3483232

[26] Nagidi, J. (2020). Best ways to handle imbalanced data in machine learning. Dataaspirant Homepage. Dataaspirant. https://dataaspirant.com/handle-imbalanced-data-machine-learning/.

[27] Shi, S., Li, J., Zhu, D., Yang, F., Xu, Y. (2023). A hybrid imbalanced classification model based on data density. Information Sciences, 624: 50-67. https://doi.org/10.1016/j.ins.2022.12.046

[28] Tallo, T.E., Musdholifah, A. (2018). The implementation of genetic algorithm in smote (synthetic minority oversampling technique) for handling imbalanced dataset problem. In 2018 4th International Conference on Science and Technology (ICST), Yogyakarta, Indonesia, pp. 1-4. https://doi.org/10.1109/ICSTC.2018.8528591

[29] Tarekegn, A.N., Giacobini, M., Michalak, K. (2021). A review of methods for imbalanced multi-label classification. Pattern Recognition, 118: 107965. https://doi.org/10.1016/j.patcog.2021.107965

[30] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16: 321-357. https://doi.org/10.1613/jair.953

[31] He, H., Bai, Y., Garcia, E.A., Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, pp. 1322-1328. https://doi.org/10.1109/IJCNN.2008.4633969

[32] Brandt, J., Lanzén, E. (2021). A comparative review of SMOTE and ADASYN in imbalanced data classification. https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-432162.

[33] Sharma, S., Gosain, A., Jain, S. (2022). A review of the oversampling techniques in class imbalance problem. In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021, pp. 459-472. https://doi.org/10.1007/978-981-16-2594-7_38

[34] Novaliendry, D., Permana, A., Dwiyani, N., Ardi, N., Yang, C.H., Saragih, F.M. (2024). Development of a semantic text classification mobile application using TensorFlow Lite and Firebase ML Kit. Journal Européen des Systèmes Automatisés, 57(6): 1603-1611. https://doi.org/10.18280/jesa.570607

[35] Novaliendry, D., Ardi, N., Saari, E.M.B., Dwiyani, N. (2023). Model development of android-based learning in vocational high school. International Journal of Interactive Mobile Technologies, 17(22): 152-159. https://doi.org/10.3991/IJIM.V17I22.45403

[36] Novaliendry, D., Huda, A., Latifahannisa, R.R.K., Costa, R.R.K., Yudhistira, Eliza, F. (2023). The effectiveness of web-based mobile learning for mobile subjects on computers and basic networks in vocational high schools. International Journal of Interactive Mobile Technologies (iJIM), 17(9): 20-30. https://doi.org/10.3991/ijim.v17i09.39337

[37] Ahmad, S.R., Insani, N., Salim, M. (2024). Analysis of cyberbullying on social media using a comparison of naïve bayes, random forest, and SVM algorithms. Jurnal Teknologi Informasi dan Pendidikan, 17(1): 75-86. https://doi.org/10.24036/jtip.v17i1.807

[38] Al-aziz, H.Y., Monalisa, S. (2023). Comparison of facebook and instagram to assess the efectiveness of advertising channels in customer acquisition. Jurnal Teknologi Informasi dan Pendidikan, 15(3): 64-72. https://doi.org/10.24036/jtip.v15i2.677