International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Designing an E-Repository of Sentiment Data and Cyberbullying Detection in Indonesian using a Parameter Optimization Algorithm for LSTM

Michael Abhinaya Bagioyuwono[1] , Dinar Ajeng Kristiyanti[2*], Antonius Sony Eko Nugroho[3]

Submitted: 29/01/2024 **Revised**: 07/03/2024 **Accepted**: 15/03/2024

Abstract: The rise in the number of social media users, particularly in Indonesia, has led to an increase in the prevalence of cyberbullying cases in Indonesia. One of the rising social media platform twitter is nominated as the moxt toxic platform. One approach to preventing cyberbullying on social media is to analyse a person's opinion or assessment of the sentiment or emotion expressed on the platform. Sentiment Analysis, has been shown to be an effective strategy for identifying and addressing cyberbullying. Hyperparameter tuning is crucial for enhancing the performance of deep learning models like Long Short-Term Memory (LSTM), which often struggle with optimizing parameters due to local minima. This challenge is tackled using Particle Swarm Optimization (PSO) and Salp Swarm Algorithm (SSA) for more effective tuning.. The research encompasses various stages: data scraping, preprocessing, translating, labeling, and modeling with an LSTM optimized by PSO and SSA to determine the optimal number of LSTM units. This is followed by statistical testing and evaluation. The optimal model will be utilized in the data repository website and cyberbullying classification based on user input and allows users to download and upload datasets with administrator permission. The finding shows that models LSTM that been optimised with PSO (PSO-LSTM), has the best performance between the conventional model and the SSA-LSTM model. The PSO-LSTM algorithm produces 87.43% accuracy, 41.29% loss, and 12.93 seconds execution time. Results of the website data repository design have been tested with the User Acceptance Test with the results running in accordance with the expected results.

Keywords: Cyberbullying Detection, Database Repository, Long Short-Term Memory (LSTM), Optimizing Parameters, Particle Swarm Optimization (PSO), Salp Swarm Algorithm (SSA)

1. Introduction

The prevalence of social media use across the globe, with a particular focus on Indonesia, continues to grow at an exponential rate. The more intense and extensive use of social media can lead to several problems such as potential interactions with strangers who have bad intentions. One of the problems that can occur is exposure to online aggression such as cyberbullying [1]. Cyberbullying on social media has become a pressing issue in recent years. According to survey that conducted by three diferent platfom namely Microsoft, UNICEF and U-Report Indonesia from 2777 responded 45% have experience cyberbullying and 71% of them happened in social media platform [2].

Cyberbullying is a subset of bullying. It is defined as an act or threat perpetrated through technology in an online environment, such as social media or text messaging [3]. Individuals who have been subjected to cyberbullying exhibit more severe depressive symptoms than those who have been victimized by other forms of bullying. The inability to effectively defend oneself from cyberbullying can intensify feelings of helplessness, which in turn can precipitate feelings of fear and emotional distress, ultimately leading to depressive symptoms [4]. Victims will tend to abuse alcohol and drugs to combat the feelings of cyberbullying they are experiencing. The use of these drugs and the effects of cyberbullying itself can push victims to attempt suicide or commit suicide. Suicides caused by cyberbullying on social media have given rise to the phenomenon of Cyberbullicide. Cyberbullicide is a term introduced by renowned researchers in the field of cyberbullying, which refers to specific cases where suicide is directly or indirectly the result of aggregation or cyberbullying [5].

Indonesia has a total of 21.15 million people who use social media, an increase from the previous year, when there were 16.2 million Twitter users in Indonesia [6]. Given the sheer volume of tweets posted daily, it is evident that Twitter is unable to exert control over the environment of the social media platform. Forbes and a study from SimpleTexting have identified Twitter as the most toxic microblogging platform, with an average toxicity score of 7.28. Furthermore, Twitter has been named the app with the most trolls by 38% of respondents [7]. Despite the implementation of various features designed to mitigate cyberbullying, including filters for unwanted messages from users who do not use profile photos and a feature to activate time limits to punish users who use inappropriate words, the Twitter platform remains susceptible to cyberbullying [8]. One approach to preventing cyberbullying on social media

[1,2,3] Department of Information System, Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Tangerang, Indonesia*
[2]ORCID ID: 0000-0001-7887-9842*
** Corresponding Author Email: dinar.kristiyanti@umn.ac.id*

is to analyse a person's opinion or assessment of the sentiment or emotion expressed on the platform. This method, commonly referred to as Sentiment Analysis, has been shown to be an effective strategy for identifying and addressing cyberbullying. The process of text sentiment analysis is an automated method of determining whether a text segment contains objective or opinionated content. This allows for the identification of text sentiment polarity.

Previous research on the detection and classification of cyberbullying, as well as sentiment analysis using machine learning and deep learning algorithms, yielded inconclusive results. A study comparing LSTM and GRU found that LSTM was more effective on different datasets with better performance than GRU [9]. However, in sentiment research related to the COVID-19 pandemic, LSTM demonstrated the lowest accuracy at 65%, while other deep learning methods achieved LSTM 89% [10]. In Indonesia, research on Twitter using machine learning and deep learning algorithms recorded a maximum accuracy of KNN 70% and LSTM 79% respectively [11] [12], which is still lower than similar research on hate speech on Arabic social media with LSTM 81% accuracy [10]. A study comparing Particle Swarm Optimisation (PSO) and Salp Swarm Algorithm (SSA) demonstrated the effectiveness swarm optimization potential. Both methods exhibited enhanced accuracy and reduced execution time, with PSO demonstrating superior speed and SSA demonstrating higher accuracy in machine learning [13]. The process of collecting data from Twitter social media using web scraping techniques is often impeded by obstacles such as access restrictions imposed by platform policies that limit the number of tweets that can be viewed in a given day. To overcome this challenge, it is necessary to create an electronic repository that serves as a data storage base and supports the cyberbullying data collection process.

The contributions of this research are as follows: 1) The performance of sentiment analysis on cyberbullying words on Twitter social media with LSTM, PSO-LSTM, and SSA-LSTM algorithms in order to identify the optimal algorithm. 2) Using the most recent Tweets relevant to cyberbullying with Indonesian language from platform X in the period 31 december 2023 - 31 January 2024. 3) The creation of an E-Repository as a foundation for the storage of datasets to be utilized in research. 4) The development of a straightforward web-based application that will automatically classify tweets containing cyberbullying text.

2. Related Works

S. Yang, X. Yu, and Y. Zhou investigate the efficacy of different neural networks with varying data sizes and lengths using the Yelp review dataset [22]. The algorithm evaluated the performance of LSTM and GRU in four different configurations: long text data and small datasets, long text and large datasets, short text and small datasets, and short text and large datasets [11]. The LSTM model demonstrated superior accuracy, achieving 75% to 79% in the four datasets tested. However, the GRU model exhibited a significantly faster processing time, with an average of 129 to 130 seconds. Another study was conducted to investigate the sentiment analysis of tweets related to the novel coronavirus disease (COVID-19) using a deep learning model [10]. The research compared four deep learning models: Bidirectional Encoder Representations from Transformers (BERT), LR, SVM, and LSTM. The results of this model indicated that the BERT model achieved an accuracy of 89%, while the LR, SVM, and LSTM models attained accuracies of 75%, 74.75%, and 65%, respectively.

The detection of cyberbullying using sentiment analysis in the Indonesian language has also been previously investigated. A. Muzakir, H. Syaputra, and F. Panjaitan [11] discuss the detection of cyberbullying in social media, specifically Twitter, using a combination of classification algorithms, namely NB, DT, LR, and SVM, which are integrated with the extraction of bigrams, unigrams, and trigrams. The obtained accuracy results are relatively high, at 76% using SVM. Furthermore, the detection of cyberbullying in Indonesian language has been investigated using deep learning and hybrid deep learning approaches [12]. The algorithms employed were LSTM, CNN, LSTM-CNN, and CNN-LSTM. This research makes a novel contribution by integrating deep learning and hybrid deep learning models with word embedding techniques, namely word2vec and TF-IDF, to detect cyberbullying in Indonesian language. The highest accuracy of 79.48% was obtained by the CNN-LSTM model. The result of the CNN-LSTM model exhibited an increase of 0.48% in accuracy compared to the LSTM baseline, which achieved an accuracy of 79.10%. While this increase is not statistically significant, it does demonstrate the ability of the algorithm to detect instances of cyberbullying.

Previous research on the integration of swarm algorithms with optimisation algorithms has been done before, but in the context of machine learning. A research to find sentiment analysis on twitter dataset using swarm algorithm to increase machine learning performance has been researched before [14]. The swarm algorithms employed were the Salp Swarm Algorithm Transfer Function (SSA-TF), Salp Swarm Algorithm (SSA), Particle Swarm Optimization (PSO), and Ant Lion Optimisation (ALO), with the objective of enhancing the performance of K-Nearest Neighbour (KNN), Support Vector Machine (SVM), and Naïve Bayes (NB). The results of this research demonstrated that the SSA-V3-TF algorithm exhibited the highest accuracy, achieving an accuracy of 80.9%. Conversely, the SSA-KNN algorithm exhibited the lowest accuracy, with an accuracy of 55%. However, when compared to other algorithms without any further optimisation of the transfer function, the PSO and ALO algorithms demonstrated promising accuracy, with the KNN-PSO algorithm achieving an accuracy of 80.95% and the KNN-ALO algorithm achieving an accuracy of 80.44%. A further study was conducted to predict land and fire data from Twitter sentiment analysis using a swarm algorithm [13]. This research integrated classification algorithms NB, SVM, and K-NN with a swarm algorithm optimiser, PSO, SSA, and ALO. The results of this research indicated that, overall, KNN optimised with SSA exhibited the highest accuracy, at 90.02%. However, the most efficient execution time was achieved by combining KNN with PSO, with a time of 0.038 seconds.

The majority of previous studies did not compare the performance of the Particle Swarm Optimizaton (PSO) algorithm and the Salp Swarm Algortihm (SSA) algorithm to find the best optimisation algorithm for the Long Short-Term Memory (LSTM) deep learning algorithm in the classification of Indonesian-language cyberbullying on twitter social media. To fill the gap of research conducted previously, this research will adopt Swarm Intelligence algorithm optimisation using LSTM deeplearning algorithm and compare it with and without Swarm Intelligence based hyperparameter optimisation, the algorithm will compare 2 optimisation namely Particle Swarm Optimization (PSO) and Salp Swarm Algortigm (SSA) to improve LSTM performance in sentiment classification on cyberbullying data.

The model with the best accuracy will build a website prototype to be able to predict cyberbullying text. In addition, an E-Repository database will be developed for the datasets used in the research. The creation of this E-repository is based on two previous studies that made an E-repository for collecting grain data [15] and collecting cancer cell images [16] because there is no dedicated database for the datasets used in the research. Therefore, this research will also create a Cyberbullying E-repository as a cyberbullying text database that will be used for research.

3. Proposed Method

The propposed methodology for this research can be shown in figure 1.

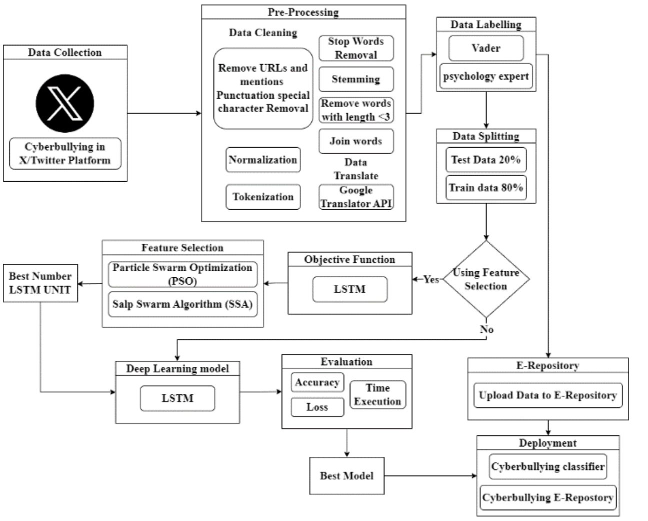


Figure 1. Proposed Architecture

3.1. Data Collection

The data to be employed in this research will be derived from primary sources, specifically from the social media platform X (Twitter). The researchers will collect this data through the use of automated data-gathering tools, also known as tweet-harvest. These tools will extract user tweets from the platform.

The research will be conducted using keywords related to cyberbullying in Indonesian, including "bajingan, anjing, tai, jablay, and goblok", as well as several Twitter threads that contain phenomena that generate various diverse opinions. The data will be gathered over the period 31 December 2023 to 31 January 2024, with a data scraping limit of 2000 data. the results of the data scraping process can be seen in the table 1.

Table 1. Tweets data cyberbullying related from Twitter/X

Created At	Full-Text
Sun Dec 31 23:58:55 +0000 2023	@laliceland yang tai tai aja
Sun Dec 31 23:42:24 +0000 2023	pgn idup normal kek org2 anjing kapan yaaahh
Sun Dec 31 23:48:07 +0000 2023	@shan_er96 lah sama banh, pas siang kaga kenapa-napa menjelang tidur udah kek anjing menggonggong kafilah berlalu
Sun Dec 31 23:58:29 +0000 2023	@convomf Nguji apaan anjing. Bilang aja antara dia pelit atau emg miskin. Najis bgt cowokmu. Putusin aja napasih

3.2. Data Pre-processing

Data pre-processing in Twitter is a transformation process to transform data so that it can be processed by the algorithms used in the research. Data pre-processing will help the classification process to identify sentiment and false information in the text, as well as reduce noise, which is very influential in determining the results of the classification algorithm used. Each Twitter user has their own punctuation style and content such as urls, hashtags and mentions that can influence the research [17]. Based on previous studies applying preprocessing to Twitter/X data, data cleaning typically involves several steps [18] [19].

3.2.1. Remove URL mention hashtag Punctuation

Unwanted strings or unicode words such as URLs, user mentions, and hastag sybmbols are categorized as leftover data that exists from the data crawling process, this leftover data can contribute to noise in dataset and does not have much value for the analytical task

3.2.2. Tokenization

Tokenization is the stage where boundaries in the text are identified such as spaces and punctuation marks. This process will break the sentence into small meaningful parts to identify the word entities in it

3.2.3. Normalization

Normalisation is the process of converting text into a standard format with the objective of facilitating consistent processing. This process can assist in reducing variation in text, thereby enhancing the ability of natural language processing algorithms and machine learning models to process it. For instance, normalisation can result in the transformation of “Utk” to “untuk” and “menunda” to “mengundurkan”.

3.2.4. StopWords Removal

Stop Word removal is the process of removing function words that cannot provide information in the index or words that usually has less meanings and does not containt any sentiment but apperar frequently in text

3.2.5. Stemming

Stemming is a process in pre-processing to remove prefixes and suffixes to reduce words that have affixes. reducing words to their root form, or "stem," by removing affixes like prefixes or suffixes. This technique helps normalize words so that variations of a word, such as “bermain” dan “mainan” will be converted base word, “main”.

3.2.6. Remove Words with length <2

After going through all the data pre-processing steps, we can compare the tweet data before and adter clensing stage. The comparison is shown on Table 2.

Table 2. Comparison of tweet data before and after pre-processing

Before Pre-Processing	After Pre-Processing
Anjing kepikiran pake AI dibikin ginian lagi 😭😭	anjing kepikiran pake dibikin ginian lagi
@miteaka Soft banget ya elu ke gue, manggil kakak sebelum anjing babian□	soft banget elu gue manggil kakak sebelum anjing babian
@lhayesno @kurawa @PreciosaKanti dia mah pasti merasa bodoamat soalnya dah tajir melintir dia dari hasil ngebuzzer😅, makanya jadi orang goblok gitu🤪🤪	dia mah pasti merasa bodoamat soalnya dah tajir melintir dia dari hasil ngebuzzer makanya jadi orang goblok gitu

3.3. Data translation

Once the data has been cleansed, the results of the cleansing process for data that is still in Indonesian will be translated into English using the Google Cloud Translation Hub API. The Google Cloud Translation API is a service provided by Google Cloud Platform that allows users to translate text from one language to another in large quantities [20]. It can translate text into higher quality. The Google Cloud Translation API allows users to translate more than 1000 languages, which can be translated in various contexts and multilingual environments. It is efficient in performing large-scale text translations. The result of data translation dan be seen in Table 3.

Table 3. Comparison of tweet data before and data translation

Before Translation	After Translation
gua nya anjing klo temen gua beda lagi asu	I'm a dog, but my friend is different and affectionate
gue kesel gue gabisa silent treatment gue sedih juga kalo ngediemin tapi gimana lagi gue sakit hati anjing	I'm annoyed that I can't get silent treatment. I'm also sad if I'm silent, but what else can I do, I'm hurt by the dog
pagipagi bukannya menghirup udara segar malah menghirup asap rokok tai	In the morning, instead of breathing fresh air, I inhaled cigarette smoke

3.4. Data Labeling

The English-language tweets data will be subjected to a sentiment score weighting process using the Valence Aware Dictionary and Sentiment Reasoner (VADER) tool.

VADER Analyzer will assign a score to the entire text, with a positive value indicating a score greater than 0 and a negative value indicating a score less than 0 [21]. The sentiment score will only store positive and negative labels derived from the compound score results determined by VADER. Once the researchers have obtained the results of the VADER labelling, these will be subjected to a second round of expert scrutiny. This will involve an examination of the labelling in question to determine whether it aligns with the text in question, which deals with the phenomenon of cyberbullying. The expert tasked with this second round of scrutiny is Rahadian Hogantara, an alumnus of the psychology programme at the State University of Jakarta (UNJ). Table 4. Represent a sample of labeled data using the vader libaray

Table 4. Result of labeled data using VADER and expert

Text	**Vader**	**Expert**
anjing lee heeseung ganteng banget	Positive	Negative
anjing tidak pantas sama bidadari	Positive	Positive
iya sudah gembok bodoh amat sudah gue kasih tantangan terbuka	Negative	Negative

3.5. Data Splitting

The data is split using the train_test_split function into two distinct sets: training data and test data. The data is split in a ratio of 20% test data and 80% training data. The X_train and X_test data have been transformed into three-dimensional tensor form, and the target data y_train and y_test have been converted into one-hot encoding in accordance with the data format requirements for running the LSTM model.

3.6. Swarm Intelligence

Swarm Intelligence algorithms are inspired by various populations of biological organisms that are made to mimic the characteristics or properties of organisms when interacting individually or in the environment to achieve a certain goal, with the aim that this interaction can be processed into an algorithm to find an optimal or near optimal solution in a heuristic way within a reasonable period of time [63. Swarm intelligence will perform optimisation to perform hyperparameter optimisation for the number of LSTM unit created as inputs. Objective funtion will be used to evaluate the performance of the LSTM model during training and provide feedback to the model on how well it predicts the training data. From the optimisation results, the most optimal number of LSTM units used for LSTM [22] that will be rebuilt using LSTM and evaluated and compared for accuracy, loss, and execution time to get the best results.

3.6.1. Particle Swarm Optimization

Particle Swarm Optimization (PSO) is an optimization algorithm inspired by the collective behavior of animal swarms in nature, particularly birds and fish, in their search for food. While searching for food birds will update their position according to their own best position and the best position within the flock resulting in an optimal formation [23]. Particle Swarm Optimisation (PSO) is an algorithm where particles in a flock, analogous to birds, move in an N-dimensional search space. Each particle has a velocity and position that represents its movement. They store information about their personal best position (Pbest) and global best position (Gbest), which enables dynamic adaptation. By adjusting the velocity based on the history of optimal positions, PSO leads to better solutions with interparticle collaboration. This process continues until a stopping condition is reached, with the goal of achieving a balance between exploration and exploitation of the optimal solution. PSO is an efficient method for finding optimal or near-optimal solution [24].

$$\begin{Bmatrix} \vec{v}_i \Leftarrow \vec{v}_i + \vec{U}(0,) \otimes (\vec{p}_i - \vec{x}_i) + \vec{U}(0,\phi_2) \otimes \left(\vec{p}_g - \vec{x}_i\right) \\ \vec{x}_i \Leftarrow \vec{x}i + \vec{v}_i \end{Bmatrix} \tag{1}$$

In the equation (1) Each particle adjusts its velocity ($\vec{v}_i$) based on its personal ($\vec{p}_i$) and global ($\vec{p}_g$) best positions by introducing random factors $\vec{U}(0,p)$ and $\vec{U}(0,g)$ that enable extensive exploration. Then, the particle positions are updated by adding the new velocity to the current position, thus facilitating the movement of particles in a better direction based on personal and collective experience in the swarm. The pseudocode of PSO algorithm can been seen in table 6 [25].

Table 5. Pseudocode of Particle Swarm Optimization

(1)	Create an array of particles, each with randomly initialized positions and velocities across D dimensions.
(2)	Begin loop:
(3)	Evaluate the fitness function for each particle in D-dimensional space.
(4)	Compare each particle's current fitness with its personal best (pbesti). If the current fitness is better: Update pbesti to the current fitness value. Update the personal position record (pi) to the current position (xi).
(5)	Identify the particle with the best performance in the neighborhood and assign its index to the variable g.

(6)	Update each particle's velocity and position using equations (1)
(7)	End the loop when a termination criterion is met, which could be achieving a desired fitness level or reaching a specified number of iterations.

3.6.2. Salp Swarm Algorithm

Salp Swarm Algorithm (SSA) is an algorithm inspired by salps and the movement of salpidae population chains in the sea in search of food proposed by S. Mirjalili in 2017 [26]. Salp chains have a leader who has optimal judgement of the environment and is often at the top of the food chain, however the leader will not directly influence the movement of the group but will influence the movement of neighbouring salps [27]. the Salp Swarm Algorithm (SSA) mimics the behaviour of a salp chain in search of food. Each chain is divided into two populations: leaders and followers. The leader, at the front of the chain, updates its position by considering the location of the food source (F), upper bound (ub), and lower bound (lb), as well as random factors (p1, p2, p3) to determine the direction and magnitude of the move. Each iteration, the leader moves the followers by taking the average of their positions and the adjacent salp. This process goes on continuously, where followers adjust their positions based on the leader, searching for the global optimum position in the n-dimensional search space [28]. The explanation of how salp algorithm work to find the best optimal solution using leader and position update can be seen in equation (2), (3), (4), and 5 [29].

$$x_j^1 = \begin{array}{l} F_j + c_1((ub_j - lb_j)c_2 + lb_j)c_3 \geqslant 0 \\ F_j - c_1((ub_j - lb_j)c_2 + lb_j)c_3 < 0 \end{array} \qquad (2)$$

$$c_1 = 2e - \left(\frac{4L}{L}\right)^2 \qquad (3)$$

$$x_j^i = \frac{1}{2}at^2 + v_0 t \qquad (4)$$

$$x_j^i = \frac{1}{2}\left(x_j^i + X_j^{i-1}\right) \qquad (5)$$

The first two equations (1) and (2) describe the position updating mechanism of the salp in search of the optimal position, where F_j can be interpreted as the target position or optimal value, while c_1 and c_2 are parameters that govern the position adaptation based on the upper bound ub_j and lower bound lb_j). Equation (3) indicates the influence of the initial acceleration and velocity in the salp movement, while equation (4) may be used to calculate the average position value between two salps, assisting in the simulation of coordinative salp movement and efficient in finding the optimal asolution. The pseudocode of SSA algorithm can been seen in table 6 [29].

Table 6. Pseudocode of Salp Swarm Algorithm

(1)	Initialize the salp population (xi with i=1, 2, 3, ..., n) with upper bound (ub) and lower bound (lb) values.
(2)	While the last criterion is not met, do: Compute the fitness value of each search agent (salp) using the value F obtained from the top-tier search engine. Modify the value of c1 using Equation (3).
(3)	For each salp: If the salp is the leading salp, do: Update the position of the leading salp using Equation (2). If the salp is a follower salp, do: Update the position of the follower salp using Equation (5).
(4)	Modify the positions of the salps based on the upper and lower bounds of the variables
(5)	Evaluate the termination criterion, if met, exit the loop.
(6)	Output the value of F as the global best solution.

3.7. Long Short-Term Memory

Long Short-Term Memory Network (LSTM) is an algorithm proposed in 1997 by Hochreiter and Schmidhuber to overcome the problem of gradient loss and gradient explosion by introducing a gated function mechanism [30]. LSTM allows for the storage of multiple temporal dependencies by changing the state vectors of cells that are deployed to capture long-term dependencies [31].. The explantion of forget gate (1), input gate (2), Candidate Cell State (3), Update Cell State (4), output gate (5), output (6) can be seen by the equation below.

$$f_t = \sigma\left(W_f \cdot [x_t, h_{t-1}] + b_f\right) \qquad (6)$$

$$i_t = \sigma(W_i \cdot [x_t, h_{t-1}] + b_i) \qquad (7)$$

$$\tilde{c_t} = tanh(W_c \cdot [x_t, h_{t-1}] + b_c) \qquad (8)$$

$$C_t = f_i \cdot C_{t-1} + i_t \cdot \tilde{C}_t \qquad (9)$$

$$O_t = \sigma(W_0 \cdot [x_t, h_{t-1}] + b_o) \qquad (10)$$

$$h_t = O_t \cdot \tanh(c_t) \qquad (11)$$

The Forget Gate (f_t in an LSTM model determines how much previous memory to retain, while the Input Gate i_t controls the influx of new information. The Candidate Cell State uses the tanh function to generate potential new values, which are then integrated into the memory cell by the Update Cell State ($\widetilde{c_t}$) using the input gate. The Output Gate (O_t) regulates the amount of information transferred to output cells, allowing the model to retain important information and discard the irrelevant. The output h_t of the LSTM represents the information processed, crucial for feature representation and decision-making in subsequent layers.

3.8. Statitiscal Test

Statistical tests are used in research to evaluate hypotheses, assess relationships between variables, and provide quantitative evidence to support conclusions, thereby enhancing the credibility and reliability of findings.

3.8.1. Shapiro Wilk Test

The Shapiro-Wilk test is one of the most popular statistical tests used to test the normality of data. The Shapiro-Wilk test is based on the value of the standardised normal expected value and the sample mean [32].The equation of shapiro wilk test can be found on eqution (12)

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad (12)$$

The Shapiro-Wilk statistic, designated by the symbol W, is employed to assess the normality of a distribution. A value of W close to 1 indicates that the sample distribution is close to normal. The number of observations in the sample, designated by the symbol n, is also considered. The coefficient a_i, determined by the order and number of n data in the sample, is obtained from the distribution table to weight the observations. The individual observations in the sample, designated by the symbol x_i, and the mean of the sample, designated by the symbol $\bar{x}$, are also considered

3.8.2. ANOVA

ANOVA is a statistical test method used to compare the means of three or more different groups to determine if there is significance among the groups. ANOVA is typically used when there is one dependent variable and one or more independent variables in the dataset. The equation of ANOVA can be found on equation [33] (13)

$$F = \frac{MS_{between}}{MS_{\omega ithin}} \qquad (13)$$

In analysis of variance (ANOVA), the F test value, D, measures differences between groups using Mean Square Between (MSB) and Mean Square Within. MSB calculates the variability between groups based on their degrees of freedom, while Mean Square Within calculates the variability within groups. Comparison of the ratio of MSB to Mean Square Within determines the statistical significance of the difference.

3.9. Hyper-Parameter Setting

The classification model utilized a deep learning algorithm known as Long Short-Term Memory (LSTM). Three models will be compared in this research, namely LSTM, LSTM using PSO, and LSTM using SSA. To ensure a fair comparison between three model, the hyperparameter for each mode will be adjust accordingly. Table 7 shower the hyperparameter setting for each model based on corresponding algorithm

Table 7. Hyperparameter settings

Algorithm	Parameters	Values
LSTM	LSTM Unit	optimized by swarm
	dropout	0.2
	Reccurent dropout	0.2
	Dense activitaion	Sigmoid
	optimizer	adam
	lb	16
	up	256
PSO	Swarm Size, max itter	15, 50
SSA	num_salps, max_iter	15,50

In Table 1, the LSTM unit will be optimized by PSO and SSA in order to identify the optimal LSTM unit for the model. In the search for the optimal unit, the upper limit is set at 256, while the lower limit is fixed at 16. In the case of the LSTM model, the initial LSTM unit to be employed will be 256. The model will then proceed to compare the performance of the algorithm in the LSTM unit, in terms of accuracy, loss, and execution time.

4. Result and discussion

4.1. LSTM Optimized using PSO and SSA

The text classification process employs two distinct methodologies: text classification utilising conventional LSTMs and text classification employing LSTMs optimised with swarm intelligence-based feature optimisation. In order to provide a more accurate comparison of the models, the researcher employs a number of metrics, including the number of LSTM units used for classification during data training, accuracy as a measure of the model's predictive accuracy on data not used for training, loss as a measure of the model's experience of loss during data prediction, and time execution as a measure of the model's completion time.

Table 8. Comparison of sentiment classification model

Model	**LSTM Unit**	**Accuracy (%)**	**Test *Loss***	**Training time**
LSTM	256	86.92	45.37	13.363725
PSO-LSTM	108	87.43	41.29	12.933906
SSA-LSTM	50	86.41	42.04	11.37750

Based on the results of testing and evaluation conducted in the research, the accuracy, loss, and execution time of the model were found as follows: LSTM algorithm produces 86.92% accuracy, 45.37% loss, and 13.36 seconds execution time. PSO-LSTM algorithm produces 87.43% accuracy, 41.29% loss, and 12.93 seconds execution time. SSA-LSTM algorithm produces 86.41% accuracy, 42.04% loss, and 11.37 seconds execution time. These results show that in the implementation of LSTM for number of_lstm_unit optimisation using PSO and SSA swarm intelligence algorithms, the PSO-LSTM algorithm is the best among the three model performance results. This is because the PSO-LSTM algorithm gets an increase in accuracy of 87.43% and reduces the loss to the lowest level at 41.29% with an average execution time of 12.93 seconds. The PSO-LSTM algorithm managed to increase the accuracy by 0.51%, reduce the loss by 4.08%, and speed up the execution time by 0.43 milliseconds which makes this algorithm has the best performance among other models.

From the training of the PSO-LSTM model, the accuracy and loss graphs can be seen in Figure 2. From the visualisation of the accuracy graph, it can be seen that the training accuracy curve (blue) shows a steady increase as the number of training epochs increases. However, there are larger fluctuations in the validation accuracy curve in the final stage of training, indicating the possibility of overfitting in the PSO-LSTM model. Meanwhile, on the loss graph, the training loss curve (blue) shows a consistent decrease, indicating an attempt to minimise the loss. However, the validation loss curve shows a decreasing trend up to a point, but starts to increase again, indicating overfitting of the model.

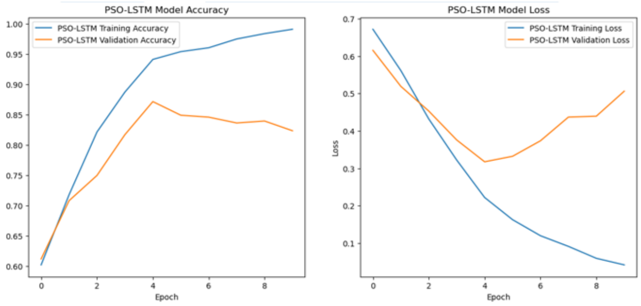


Fig 2. Model Accuracy and loss using PSO-LSTM algorithm

Figure 3. is the result of the confusion matrix, which shows the performance of the LSTM model optimised to find the best number of LSTM units using Particle Swarm Optimization? The PSO-LSTM model successfully classifies the cyberbullying sentences into 2 conditions, positive and negative. The PSO-LSTM model successfully predicted 222 words as negative sentences, but the model made an error by predicting 14 sentences as positive. For the positive condition, the model successfully predicted 119 sentences as positive sentences. However, there is a prediction error of 35 sentences predicted as negative. To prove that the PSO-LSTM model is the best model, a statistical test will be carried out to help ensure that the data results do not appear by chance and are not biased and there is a significance of the model created.

4.2. Statitiscal test

Statistical tests will be conducted using Shapiro and Anova. Shapiro test will be used to test the normality of the model and t-test is used to test the significance between models. Table 9. 4 shows the results of the Shapiro test and anova test for LSTM, PSO-LSTM, SSA-LSTM.

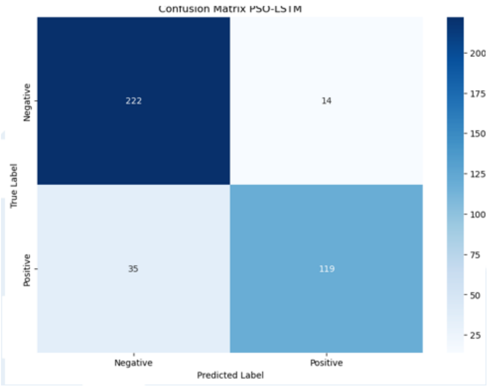


Fig 3. Confusion Matrix of LSTM using PSO-LSTM Algorithm

Table 9. Statitiscal test result using shapiro wilk and anova

Statiscal Test	LSTM Optimizer		
	LSTM	PSO-LSTM	SSA-LSTM
Shapiro-	Stat = 0.8516,	Stat =0.8396,	Stat = 0.8225,

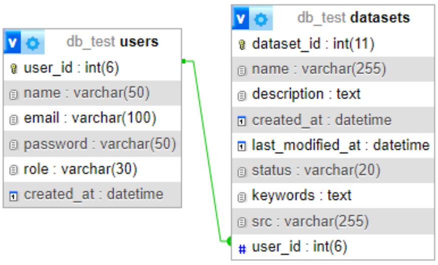
Wilk	P=0.8516	P=0.1861	P=0.0271
Anova	There are significant differences (stat=4.155, p=0.043)		

The Shapiro LSTM results show a statistical value of 0.8516 and a p-value of 0.0607 which indicates that the data is quite similar to the normal distribution and does not save from normality. Shapiro PSO-LSTM results show a statistical value of 0.8396 and a p-value of 0.1861 which indicates that the data is quite similar to normal distributions and does not save from normality. Figure 4. Shows the Shapiro test results for LSTM, PSO-LSTM, SSA-LSTM. SSA-LSTM shapiro results show a statistical value of 0.8225 and p-value of 0.0271 which indicates that the data is quite similar to normal distribution and does not save from normality. The results of the ANOVA test show that there is a significance of the compared models, the F statistic value is 4.155 and the p-value is 0.043, which indicates that the PSO-optimised LSTM model (PSO-LSTM) is proven to improve accuracy compared to the conventional LSTM model and compared to the SSA-LSTM model in data related to Indonesian cyberbullying.

4.3. Data repository and cyberbullying classification

4.3.1. Database

Figure 3 is Entity Relationship Diagram (ERD) that contains the structure of the user table that holds user data and credentials. This table will hold the primarykey user id of the user, user name, user password, and role which will be identified by numbers to define the user's role, as well as to set the feature restrictions of the user. This table is an important table because it is needed by other tables in the operation of the application such as login and signup, the user table is also used to distinguish the role of user and admin in the application website. Another table in the ERD is datasets table which holds all information and dataset files uploaded by the user. This table will store the primarykey dataset_id which will store the id of each uploaded dataset, user_id foreign key which will store the user id that uploads the dataset, dataset_name, dataset_desc, dataset_uploadate, dataset_keyword dataset_upload, status which identifies the status of the uploaded dataset whether pending approve or reject. Website ERD can be seen on Figure 4.



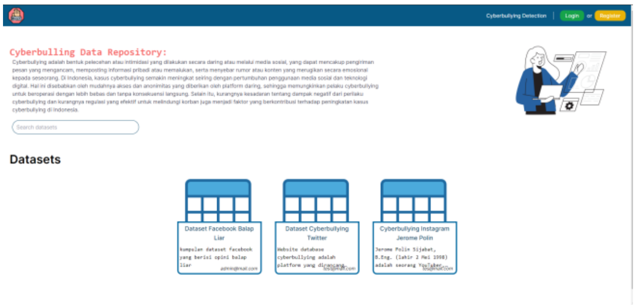
db_test users
- user_id : int(6)
- name : varchar(50)
- email : varchar(100)
- password : varchar(50)
- role : varchar(30)
- created_at : datetime

db_test datasets
- dataset_id : int(11)
- name : varchar(255)
- description : text
- created_at : datetime
- last_modified_at : datetime
- status : varchar(20)
- keywords : text
- src : varchar(255)
- user_id : int(6)

Fig 4. ERD Website

4.3.2. Website

The system created is a website repository system to create a database to store data related to cyberbullying which allows users to download and upload from the E-repository. The system also allows the admin to review the dataset before the data is uploaded to the user. The first website prototype that will be build is from the user side. Figure 5 is showing display of the user home page when the user first opens the e-repository website such as displays information and the purpose of creating a website, displays a search bar to perform a search based on the user's input. The home page also acts as a preview page to show a list of the records contained in the site. Each card will display the title, initial description of the dataset, and email of the dataset owner. If the user wants to see a more detailed page, the user can press the card to move to the dataset detail page.



Cyberbulling Data Repository:

Cyberbullying adalah bentuk pelecehan atau intimidasi yang dilakukan secara daring atau melalui media sosial, yang dapat mencakup pengiriman pesan yang mengancam, memposting informasi pribadi atau memalukan, serta menyebar rumor atau konten yang merugikan secara emosional kepada seseorang. Di Indonesia, kasus cyberbullying semakin meningkat seiring dengan pertumbuhan penggunaan media sosial dan teknologi digital. Hal ini disebabkan oleh mudahnya akses dan anonimitas yang diberikan oleh platform daring, sehingga memungkinkan pelaku cyberbullying untuk beroperasi dengan lebih bebas dan tanpa konsekuensi langsung. Selain itu, kurangnya kesadaran tentang dampak negatif dari perilaku cyberbullying dan kurangnya regulasi yang efektif untuk melindungi korban juga menjadi faktor yang berkontribusi terhadap peningkatan kasus cyberbullying di Indonesia.

Search datasets

Datasets

Dataset Facebook Balap Liar
kumpulan dataset facebook yang berisi opini balap liar
admin@mail.com

Dataset Cyberbullying Twitter
Website database cyberbullying adalah platform yang dirancang...
test@mail.com

Cyberbullying Instagram Jerome Polin
Jerome Polin Sijabat, B.Eng. (lahir 2 Mei 1998) adalah seorang YouTuber...
test@mail.com

Fig 5. Home Page of the website

Figure 6 is displaying user interface when user want to login or register. Upon accessing the register page via the navigation bar, users are presented with a series of input forms that must be completed in order to register. These include fields for user identity, such as name, email address, and password. All of these fields must be populated by the user. Once all the required information has been entered, the register button becomes active, allowing the user to proceed with the registration process. Once an account has been registered, the user can log in by entering their email address and password. These details are checked against the user database to ensure they are valid.

Figure 7. is displaying the user interface when user pressed one of the dataset cards either on the home page or the user upload dashboardon this page the information displayed on the dataset in more detail. Users can see the title, the date the dataset was uploaded, the user who uploaded it, the database search keywords, and a preview of the first 10 columns of the dataset. Users who can download the dataset are users who have logged into the website by pressing the download button at the bottom right.

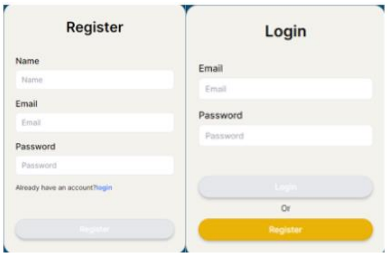
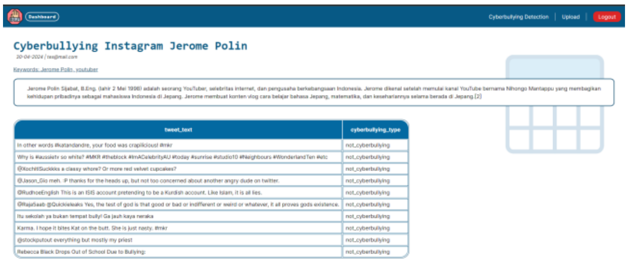


Fig 6. page login and register



Cyberbullying Instagram Jerome Polin

20-04-2024 / tes@mail.com

Keywords: Jerome Polin, youtuber

Jerome Polin Sijabat, B.Eng. (lahir 2 Mei 1998) adalah seorang YouTuber, selebritas internet, dan pengusaha berkebangsaan Indonesia. Jerome dikenal setelah memulai kanal YouTube bernama Nihongo Mantappu yang membagikan kehidupan pribadinya sebagai mahasiswa Indonesia di Jepang. Jerome membuat konten vlog cara belajar bahasa Jepang, matematika, dan kesehariannya selama berada di Jepang.[2]

tweet_text	cyberbullying_type
In other words #katandandre, your food was crapilicious! #mkr	not_cyberbullying
Why is #aussietv so white? #MKR #theblock #ImACelebrityAU #today #sunrise #studio10 #Neighbours #WonderlandTen #etc	not_cyberbullying
@XochitlSuckkks a classy whore? Or more red velvet cupcakes?	not_cyberbullying
@Jason_Gio meh. :P thanks for the heads up, but not too concerned about another angry dude on twitter.	not_cyberbullying
@RudhoeEnglish This is an ISIS account pretending to be a Kurdish account. Like Islam, it is all lies.	not_cyberbullying
@RajaSaab @Quickieleaks Yes, the test of god is that good or bad or indifferent or weird or whatever, it all proves gods existence.	not_cyberbullying
Itu sekolah ya bukan tempat bully! Ga jauh kaya neraka	not_cyberbullying
Karma. I hope it bites Kat on the butt. She is just nasty. #mkr	not_cyberbullying
@stockputout everything but mostly my priest	not_cyberbullying
Rebecca Black Drops Out of School Due to Bullying:	not_cyberbullying

Fig 7. Detailed Dataset of the Website

Logged in users can access the upload page as shown in Figure 8. On this page the user can fill in all the information fields about the dataset to be uploaded, such as Title, Dataset Description which contains information about the dataset the user is uploading, Keyword to extract data from the dataset and finally the user needs to upload the .CSV dataset file in the Upload file column. Once all the information fields have been filled in, the Upload button at the bottom will be active and the user can press the button to upload all the information to the database. If the user presses the button and the upload is successful, a popup warning will appear informing the user that the data has been successfully uploaded. However, even though the dataset has been uploaded, the data will not appear on the dataset homepage immediately. This is because any dataset uploaded by the user must first be reviewed by the administrator. Data that has been uploaded by the user can be viewed on the user dashboard page along with the status of the data.

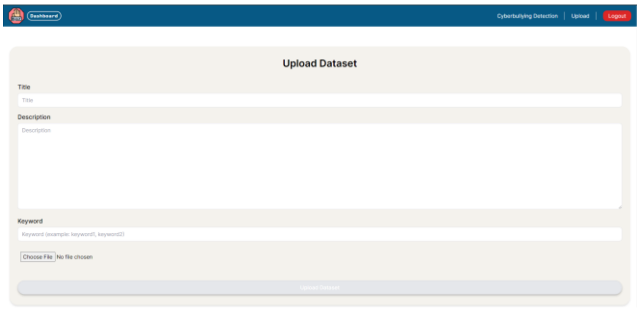
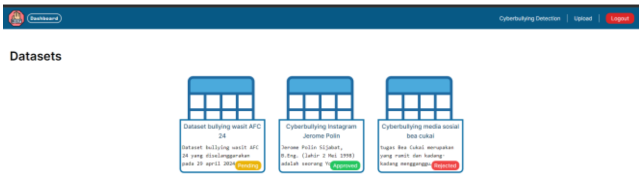


Fig 8. Upload Dataset Page

Figure 9 shows the user page when the Dashboard page is opened. On this page the user can see a list of datasets that the user has uploaded to the website and the status of the uploaded dataset. The status of the dataset has 3 states, namely Pending, Approved and Rejected, as shown in the figure. When a dataset is uploaded to the website, it is given a pending status, which means that the dataset has not been or is still awaiting the review stage by the website administrator, approved, which means that the dataset has been approved by the administrator, and rejected, which means that the dataset has not been approved by the administrator.



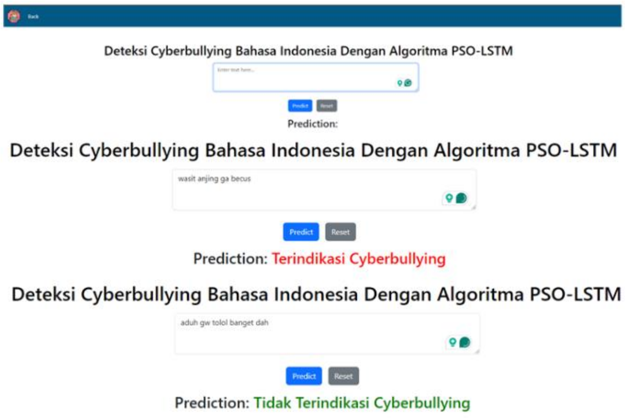
Dataset bullying wasit AFC 24
Dataset bullying wasit AFC 24 yang diselenggarakan pada 29 april 2024 Pending

Cyberbullying Instagram Jerome Polin
Jerome Polin Sijabat, B.Eng. (lahir 2 Mei 1998) adalah seorang Yo Approved

Cyberbullying media sosial bea cukai
tugas Bea Cukai merupakan yang rumit dan kadang-kadang mengganggu Rejected

Fig 9. Dashboard user

Figure 10 presents a cyberbullying detection page that users can access and utilise even in the absence of a login to the website. This page comprises a text field where the user can enter the text that they wish to be classified as cyberbullying. Upon submission, the system, connected via API with the machine learning model that has been constructed previously, will detect the text and identify whether the text contains cyberbullying or not. In the event that the text in question is deemed to contain cyberbullying, a text will appear which provides information as to the likelihood of the text in question containing cyberbullying.



Deteksi Cyberbullying Bahasa Indonesia Dengan Algoritma PSO-LSTM

Prediction:

Deteksi Cyberbullying Bahasa Indonesia Dengan Algoritma PSO-LSTM

wasit anjing ga becus

Prediction: Terindikasi Cyberbullying

Deteksi Cyberbullying Bahasa Indonesia Dengan Algoritma PSO-LSTM

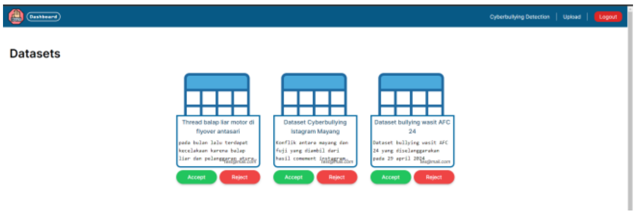
aduh gw tolol banget dah

Prediction: Tidak Terindikasi Cyberbullying

Fig 10. Upload of the Website

The next in contrast to the user account, the administrator account does not have its own dedicated account creation process. Instead, administrator accounts are created and entered directly into the database by intervening in the database. Furthermore, if the administrator role has been verified when logging in, the administrator view will be redirected to the administrator page.

Figure 11 illustrates the administrative dashboard, which can be accessed by pressing the button on the navigation bar. The dashboard displays a list of datasets uploaded by the user to the website. On this page, the administrator can edit the dataset by selecting the relevant card to preview the uploaded data and approve or reject the dataset on the card home dashboard page. Upon approval by the administrator, the status of the dataset is changed to "approved." This action renders the data visible to other users, who may then preview and download it.



Thread balap liar motor di flyover antasari

pada bulan lalu terdapat kecelakaan karena balap liar dan pelanggaran atara

Dataset Cyberbullying Istagram Mayang

Konflik antara mayang dan fuji yang diambil dari hasil comment instagram

Dataset bullying wasit AFC 24

Dataset bullying wasit AFC 24 yang diselenggarakan pada 29 april 2024

Fig 11. Dashboard Admin

4.3.3. Indonesian language cyberbullying detection testing

The subsequent stage of the study will involve the implementation of a web-based application, as depicted in Figure 4., for the purpose of detecting instances of cyberbullying. The application will be tested using a series of newly generated text samples, which have been compiled based on keywords that are commonly associated with cyberbullying on the Twitter social media platform. The Cyberbullying Prediction Testing can be seen in table 4. On the table only 5 of the 10 sample that been tested shown on table 10.

Table 10. (5) out of (10) Cyberbullying Prediction Testing samples

Sample	Target	Predicted
Mukanya kek pemeran yg dilawan karate kids, anjing lu	Positive	positive
keliatan jelas kalo wasit bayaran anjrit berat sebelah	Positive	positive
Tapi secara permainan kita memang masih dibawah uzbek bro. Kecepatan dan skill juga kita masih di bawah uzbek.	Negative	Negative
BANGSATTTTT gw trauma ANJING DENGERIN LAGU LEWAT HEADPHONE NYA BASEUSSSS ANJGG	Positive	Negative
lu tuh lebih bangsat dari pada gua	Negative	Positive

In the case of cyberbullying detection testing based on 10 sample data points extracted from Twitter, the results indicate that there are four instances of correct identification of positive targets, while there are four instances of incorrect identification of positive targets. Additionally, there are two instances of misclassification, with positive cyberbullying being identified as negative and negative cyberbullying being identified as positive. The statistical test results of data sampling from Twitter/X conducted by researchers indicate that the accuracy score of the model is 80%.

4.3.4. Blackbox Testing E-Repository

The results of user acceptance testing for the development of the E-Repository from both user and admin roles demonstrate that all features related to the E-Repository have been successfully implemented in accordance with the expected results. Consequently, all features on the website system for the user and admin side have been tested and declared successful in passing UAT. Result of Blackbox testing can be seen in Table 11.

Table 11. Blackbox testing result

Feature	User 1	User 2
Home Page	Pass	Pass
Login Register Page	Pass	Pass
Detailed dataset	Pass	Pass
Upload dataset	Pass	Pass
User dashboard	Pass	Pass
Cyberbullying Detection	Pass	Pass
Admin dashboard	Pass	

5. Discussion

The results of testing and evaluation of all models indicate that the application of LSTM unit optimisation with feature selection based on swarm intelligence can enhance the performance of the conventional LSTM algorithm. This improvement is evident in the results of the three algorithms in terms of accuracy, loss, and execution time. The LSTM algorithm achieved an accuracy of 86.92%, a loss of 45.37%, and an execution time of 13.36 seconds. The PSO-LSTM algorithm produced an accuracy of 87.43%, a loss of 41.29%, and an execution time of 12.93 seconds. The SSA-LSTM algorithm achieved an accuracy of 86.41%, a loss of 42.04%, and an execution time of 11.37 seconds. These results demonstrate that the PSO-LSTM algorithm exhibited superior performance in terms of accuracy, loss, and execution time compared to the conventional LSTM and SSA-LSTM algorithms in classifying cyberbullying on Twitter social media.

Nevertheless, this study also revealed that the three models exhibited a proclivity towards overfitting in both accuracy and loss results, as evidenced by the accuracy and loss graphs of the models. Table presents a compartive study across different research methodologies using accuracy measurement statistics in cyberullying detection in indonesian langguage. A. Muzakir, H. Syaputra, and F. Panjaitan [11] achived an accuracy of 76% using SVM, I. A. Asqolani and E. B. Setiawan [12] achieved an accuracy of 79.48% using hybrid LSTM-CNN. In Comparison, the proposed approach utilizing LSTM optimized by Particle Swarm Optimization outperform these previous researches with an accuracy of 87.43%. Table 12 provides a succinct overview of the efficacy of the proposed methodology in attaining a superior degree of accuracy in comparison to the referenced studies.

Table 12. comparison with previous research works

Judul	Algoritma	Hasil Akurasi
Proposed Method	PSO-LSTM	87.43%
A. Muzakir, H. Syaputra, and F. Panjaitan [11]	SVM	76%
I. A. Asqolani and E. B. Setiawan [12]	LSTM-CNN	79.48%

6. Conclussion

The objective of enchance cyberbullying detectionby utilising 2000 data extracted from Indonesian Twitter. These data were derived from various Twitter threads, each containing one of the following keywords: "bajingan, anjing, tai, jablay, and goblok". This research project is concerned with the development of LSTM models that have optimised with Particle Swarm Optimization (PSO) and Salp Swarm Algorithm (SSA) to reduce number of lstm unit from 256 to 108 (PSO) and 50 (SSA). The performance of each model is evaluated based on three key metrics: accuracy, loss, and execution time. The results of the testing and evaluation conducted in the study are as follows: The LSTM algorithm achieved an accuracy of 86.92%, a loss of 45.37%, and an execution time of 13.36 seconds. The PSO-LSTM algorithm attained an accuracy of 87.43%, a loss of 41.29%, and an execution time of 12.93 seconds. The SSA-LSTM algorithm achieved 86.41% accuracy, 42.04% loss, and 11.37 seconds execution time. These results demonstrate that in the implementation of LSTM for number_of_lstm_unit optimisation using PSO and SSA swarm intelligence algorithms, the PSO-LSTM algorithm outperforms the other two models in terms of performance. This is due to the fact that the PSO-LSTM algorithm exhibits an 87.43% increase in accuracy and a reduction in loss to its lowest level at 41.29% with an average execution time of 12.93 seconds. The PSO-LSTM algorithm demonstrated an increase in accuracy of 0.51%, a reduction in loss of 4.08%, and a speedup in execution time of 0.43 milliseconds. These results indicate that this algorithm has the best performance among other models.

From the algorithm performance results, this research has also designed a prototype implementation of the best PSO-LSTM modelling results to predict and classify cyberbulling text from platform x or twitter social media. In integrating the model into the prototype, the flask framework is used in developing web application pages on the E-Repository website based on the trained model that can perform classification based on input text from users who use the website. Users can enter text data in the column and the algorithm will classify whether the text contains cyberbullying or not based on the PSO-LSTM optimisation algorithm that has been developed. The cyberbullying classification prototype has the potential to classify cyberbullying on twitter social media.

The final result of the research successfully produced a prototype E-repository website that focuses as a research database that can be accessed by users and admins. The website is designed based on research needs as a database allowing users to download datasets, upload datasets, preview datasets, and classify cyberbullying with user input which is the implementation of deep learning models in research. In addition to this, in terms of the admin role, it succeeded in implementing dataset review as a security for user uploads that must pass admin review before the dataset is released to the user. This research limitation includes optimizing LSTM model hyperparameter using LSTM unit using swarm intelligence particle swarm optimization and salp swarm alogirthm using Indonesia tweets focusing on cyberbullying keywords and several twitter thread and has labeled and confirmed by expert.

6.1. Future Work

In order to enhance the efficacy of the cyberbullying classification system for future research, it is recommended that more comprehensive data scraping be conducted with a broader range of datasets, thereby enabling the system to identify a more diverse array of data. The subsequent crucial step is to prevent overfitting by implementing regularisation techniques such as dropout and L1 and L2 penalties, in addition to augmenting the model with supplementary parameters and LSTM layers to enhance its complexity. Furthermore, the utilisation of a wider range of swarm intelligence optimisation techniques, such as Grey Wolf Optimisation, Ant Lion Optimisation, and Social Spider Algorithm, in conjunction with an increasing swarm size, can assist in enhancing the optimisation capabilities. Finally, the development of an interactive website application that can be accessed in real time and supports real-time detection of diverse datasets will considerably strengthen the practical applicability of this system.

Acknowledgment

The author would like to express gratitude to the Universitas Multimedia Nusantara for funding and support this research.

Author contributions

Michael Bagioyuwono: Data Curation, Conceptualization, Methodology, Software, Formal Analysis. Validation, Writing – Original Draft Preparation. **Dinar Ajeng Kristiyanti:** Conceptualization, Supervision, Methodology, Writing – Review & Editing. **Antonius Sony Eko Nugroho:** Conceptualization, Supervision.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] W. Craig et al., “Social Media Use and Cyber-Bullying: A Cross-National Analysis of Young People in 42 Countries,” Journal of Adolescent Health, vol. 66, no. 6, 2020, doi: 10.1016/j.jadohealth.2020.03.006.

[2] S. A. Anggraeni, L. J. H. Lotulung, and ..., “Motif Perilaku Cyberbullying Remaja Di Media Sosial Twitter,” Acta Diurna …. 2022.

[3] M. H. L. Lee, M. Kaur, V. Shaker, A. Yee, R. Sham, and C. S. Siau, “Cyberbullying, Social Media Addiction and Associations with Depression, Anxiety, and Stress among Medical Students in Malaysia,” Int J Environ Res Public Health, vol. 20, no. 4, Feb. 2023, doi: 10.3390/IJERPH20043136.

[4] S. Kaiser, H. Kyrrestad, and S. Fossum, “Cyberbullying status and mental health in Norwegian adolescents,” Scand J Psychol, vol. 61, no. 5, pp. 707–713, Oct. 2020, doi: 10.1111/sjop.12656.

[5] Schonfeld, D. McNiel, T. Toyoshima, and R. Binder, “Cyberbullying and Adolescent Suicide,” Journal of the American Academy of Psychiatry and the Law Online, vol. 51, no. 1, pp. 112–119, Mar. 2023, doi: 10.29158/JAAPL.220078-22.

[6] S. Kemp, “TWITTER STATISTICS AND TRENDS,” DATE REPORTAL. [Online]. Available: https://datareportal.com/essential-twitter-stats?utm_source=DataReportal&utm_medium=Country_Article_Hyperlink&utm_campaign=Digital_2022&utm_term=Indonesia&utm_content=Facebook_Stats_Link

[7] “Twitter Tops the List of Most Toxic Apps.” Accessed: Mar. 01, 2024. [Online]. Available: https://www.forbes.com/sites/petersuciu/2022/06/08/twitter-tops-the-list-of-most-toxic-apps/?sh=6cb110565d53

[8] V. Balakrishnan, S. Khan, and H. R. Arabnia, “Improving cyberbullying detection using Twitter users’ psychological features and machine learning,” Comput Secur, vol. 90, 2020, doi: 10.1016/j.cose.2019.101710.

[9] S. Yang, X. Yu, and Y. Zhou, “LSTM and GRU Neural Network Performance Comparison Study: Taking Yelp Review Dataset as an Example,” Proceedings - 2020 International Workshop on Electronic Communication and Artificial Intelligence, IWECAI 2020, pp. 98–101, Jun. 2020, doi: 10.1109/IWECAI50956.2020.00027.

[10] N. Chintalapudi, G. Battineni, and F. Amenta, “Sentimental Analysis of COVID-19 Tweets Using Deep Learning Models,” Infectious Disease Reports 2021, Vol. 13, Pages 329-339, vol. 13, no. 2, pp. 329–339, Apr. 2021, doi: 10.3390/IDR13020032.

[11] Muzakir, H. Syaputra, and F. Panjaitan, “A Comparative Analysis of Classification Algorithms for Cyberbullying Crime Detection: An Experimental Study of Twitter Social Media in Indonesia,” Scientific Journal of Informatics, vol. 9, no. 2, pp. 133–138, Oct. 2022, doi: 10.15294/SJI.V9I2.35149.

[12] A. Asqolani and E. B. Setiawan, “A Hybrid Deep Learning Approach Leveraging Word2Vec Feature Expansion for Cyberbullying Detection in Indonesian Twitter,” Ingenierie des Systemes d’Information, vol. 28, no. 4, 2023, doi: 10.18280/isi.280410.

[13] D. A. Kristiyanti, I. S. Sitanggang, Annisa, and S. Nurdiati, “Feature Selection Technique Model for Forest and Land Fire Data Sentiment Analysis: Comparison of SSA, PSO, and ALO,” in Proceedings of the 7th 2023 International Conference on New Media Studies, CONMEDIA 2023, 2023. doi: 10.1109/CONMEDIA60526.2023.10428170.

[14] M. Sosnowski et al., “Feature Selection Using New Version of V-Shaped Transfer Function for Salp Swarm Algorithm in Sentiment Analysis,” Computation 2023, Vol. 11, Page 56, vol. 11, no. 3, p. 56, Mar. 2023, doi: 10.3390/COMPUTATION11030056.

[15] N. Govind, M. Sahoo, S. S. K. Pillai, and S. K. Sahu, “IPSD: e-repository of Permian seeds from Indian Lower Gondwana,” Acta Palaeobotanica, vol. 63, no. 2, pp. 151–161, Dec. 2023, doi: 10.35535/ACPA-2023-0010.

[16] D. Riana et al., “Identifikasi Citra Pap Smear RepoMedUNM dengan Menggunakan K-Means Clustering dan GLCM,” Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 6, no. 1, pp. 1–8, Jan. 2022, doi: 10.29207/RESTI.V6I1.3495.

[17] U. Naseem, I. Razzak, and P. W. Eklund, “A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter,” Multimedia Tools and Applications 2020 80:28, vol. 80, no. 28, pp. 35239–35266, Nov. 2020, doi: 10.1007/S11042-020-10082-6.

[18] U. Naseem, I. Razzak, and P. W. Eklund, “A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter,” Multimed Tools Appl, vol. 80, no. 28–29, pp. 35239–35266, Nov. 2021, doi: 10.1007/S11042-020-10082-6/METRICS.

[19] H.-T. Duong and A. Nguyen-Thi, “A review: preprocessing techniques and data augmentation for sentiment analysis”, doi: 10.1186/s40649-020-00080-x.

[20] H. H. Wang, “Speech Recorder and Translator using Google Cloud Speech-to-Text and Translation,” Journal of IT in Asia, vol. 9, no. 1, 2021, doi: 10.33736/jita.2815.2021.

[21] M. Suyal and P. Goyal, “A New Classifier Model on Drug Reviews Dataset by VADER Sentiment Analyzer to Analyze Reviews of the Dataset are Real or Fake based on Machine Learning,” International Journal of Engineering Trends and Technology, vol. 70, no. 7, pp. 68–78, Jul. 2022, doi: 10.14445/22315381/IJETT-V70I7P208.

[22] C. Chen, N. Lu, L. Wang, and Y. Xing, “Intelligent selection and optimization method of feature variables in fluid catalytic cracking gasoline refining process,” Comput Chem Eng, vol. 150, p. 107336, Jul. 2021, doi: 10.1016/j.compchemeng.2021.107336.

[23] M. Jain, V. Saihjpal, N. Singh, and S. B. Singh, “An Overview of Variants and Advancements of PSO Algorithm,” Applied Sciences 2022, Vol. 12, Page 8392, vol. 12, no. 17, p. 8392, Aug. 2022, doi: 10.3390/APP12178392.

[24] J. Zhou et al., “Predicting TBM penetration rate in hard rock condition: A comparative study among six XGB-based metaheuristic techniques,” Geoscience Frontiers, vol. 12, no. 3, p. 101091, May 2021, doi: 10.1016/J.GSF.2020.09.020.

[25] R. Poli, J. Kennedy, and T. Blackwell, “Particle Swarm Optimization: An Overview Particle swarm optimization an overview”, doi: 10.1007/s11721-007-0002-0.

[26] M. Sosnowski et al., “Feature Selection Using New Version of V-Shaped Transfer Function for Salp Swarm Algorithm in Sentiment Analysis,” Computation 2023, Vol. 11, Page 56, vol. 11, no. 3, p. 56, Mar. 2023, doi: 10.3390/COMPUTATION11030056.

[27] Q. Duan, L. Wang, H. Kang, Y. Shen, X. Sun, and Q. Chen, “Improved Salp Swarm Algorithm with Simulated Annealing for Solving Engineering Optimization Problems,” Symmetry 2021, Vol. 13, Page 1092, vol. 13, no. 6, p. 1092, Jun. 2021, doi: 10.3390/SYM13061092.

[28] S. Kassaymeh, S. Abdullah, M. A. Al-Betar, and M. Alweshah, “Salp swarm optimizer for modeling the software fault prediction problem,” Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 6, pp. 3365–3378, Jun. 2022, doi: 10.1016/J.JKSUCI.2021.01.015.

[29] L. P. Hung and S. Alias, “Beyond Sentiment Analysis: A Review of Recent Trends in Text Based Sentiment Analysis and Emotion Detection,” Journal of Advanced Computational Intelligence and Intelligent Informatics, vol. 27, no. 1. 2023. doi: 10.20965/jaciii. 2023.p0084.

[30] Picornell et al., “A deep learning LSTM-based approach for forecasting annual pollen curves: Olea and Urticaceae pollen types as a case study,” Comput Biol Med, vol. 168, p. 107706, Jan. 2024, doi: 10.1016/J.COMPBIOMED.2023.107706.

[31] H. Chen, X. Li, Y. Wu, L. Zuo, M. Lu, and Y. Zhou, “Compressive Strength Prediction of High-Strength Concrete Using Long Short-Term Memory and Machine Learning Algorithms,” Buildings 2022, Vol. 12, Page 302, vol. 12, no. 3, p. 302, Mar. 2022, doi: 10.3390/BUILDINGS12030302.

[32] G. Dwijuna Ahadi, N. N. Laili, and E. Zain, “The Simulation Study of Normality Test Using Kolmogorov-Smirnov, Anderson-Darling, and Shapiro-Wilk,” EIGEN MATHEMATICS JOURNAL, vol. 6, no. 1, pp. 11–19, Jun. 2023, doi: 10.29303/EMJ.V6I1.131.

[33] T. K. Kim, “Understanding one-way anova using conceptual figures,” Korean J Anesthesiol, vol. 70, no. 1, pp. 22–26, Feb. 2017, doi: 10.4097/kjae.2017.70.1.22.