

Abusive comment identification on Indonesian social media data using hybrid deep learning

Tiara Intana Sari¹, Zalfa Natania Ardilla¹, Nur Hayatin¹, Ruhaila Maskat²

¹Department of Informatics, Faculty of Engineering, University of Muhammadiyah Malang, Malang, Indonesia

²Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Shah Alam, Selangor, Malaysia

Article Info

Article history:

Received Oct 2, 2021

Revised Apr 4, 2022

Accepted Apr 28, 2022

Keywords:

Abusive comments

Deep learning

Long short-term memory

Recurrent neural network

Sentiment analysis

ABSTRACT

Half of the entire social media users in Indonesia has experienced cyberbullying. Cyberbullying is one of the treatments received as an attack with abusive words. An abusive word is a word or phrase that contained harassment and is expressed by it spoken or in the form of text. This is a serious problem that must be controlled because the act has an impact on the victim's psychology and causes trauma resulting in depression. This study proposed to identify abusive comments from social media in Indonesian language using a deep learning approach. The architecture used is a hybrid model, a combination between recurrent neural network (RNN) and long short-term memory (LSTM). RNN can map the input sequences to fixed-size vectors on hidden vector components and LSTM implemented to overcome gradient vector growth components that have the potential to exist in RNN. The steps carried out include preprocessing, modelling, implementation, and evaluation. The dataset used is Indonesian abusive and hate speech from Twitter data. The evaluation result showed that the model proposed produced an f-measure value of 94% with an increase in accuracy of 23%.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Nur Hayatin

Department of Informatics, Faculty of Engineering, University of Muhammadiyah Malang

St. Raya Tlogomas 246 Malang, Indonesia

Email: noorhayatin@umm.ac.id

1. INTRODUCTION

Almost 197 million people in Indonesia have used the internet in their daily life. This figure is directly proportional to the increase in social media users in Indonesia [1]. Indonesian society uses social media to publish their views and opinions under various circumstances [2]. One of them is through the social media used by most users to express their attitudes, thoughts or opinions on various occasions [3], [4]. A social media platform and social networking service that is widely used and generates a large amount of information are Twitter [5].

Social media has benefits for communication, marketing, and community education. On the other hand, it can cause offensive outcomes such as hate speech spreading and online harassment, as known by many with the term cyberbullying. The case of cyberbullying has become a point of criticism as well as pressure on popular social media platforms, such as Facebook and Twitter, being a problem that must be solved [6]. There are still many social media users who communicate with abusive and uncontrollable words [7]. Abusive word is one of the types of cyberbullying expressed or spoken using either orally or in text. This treatment can be an act on social media through comments. It can be a word or phrase that are rude or dirty, in the context of jokes, sex harassment, or cursing someone [8].

Komisi Perlindungan Anak Indonesia (The Indonesian Child Protection Commission) stated that there were 37381 complaints regarding bullying cases from 2011 to 2019 where there is a total of 2473 data

related to cyberbullying on social media [9]. Other data, published by the association of Indonesia internet service providers in 2019, shows that cyberbullying cases in Indonesia have reached approximately 49% [10]. These data are proof that cyberbullying is worrying and has to be controlled properly, including utilized abusive words on social media. It impacted a victim's psychology and caused trauma. Other than that, there is a tendency for the victim to experience anxiety, individualism or even antisocial behaviour, resulting in depression due to prolonged cyberbullying [1].

Nowadays, the utilization of abusive words is uncontrollable in social media particularly [8]. Detection of abuse in user-generated online content on social media is a difficult but important task [11], [12]. However, identifying abusive comments need a substantial effort if done manually. Machine learning (ML) can be used to classify comments that contained abusive words automatically. Ibrohim and Budi have studied the problem of abusive comments on Indonesian tweets. They compared several methods in machine learning such as random forest decision tree (RFDT), support vector machine (SVM), and naïve Bayes (NB). The result shows that NB produced higher accuracy than other techniques. However, it does not provide a good enough performance with accuracy is around 70.06% [13].

Deep learning (DL) is a new approach that has a good performance for a classification task which was proven from recent research. The study proposed by Chakraborty and Seddiqui has compared ML and DL approaches. They used machine learning to detect abusive context on social media in the Bengali language. This research implements SVM and multinomial naïve Bayes (MNB) as a method to detection abusive language. They also implement a deep learning method that is using a combined model convolutional neural network (CNN) with long short-term memory (LSTM). From all models that were used, the best result was achieved by the SVM method with 78% accuracy [14]. A similar approach was used in [15] to detect an abusive context on Urdu and Roman Urdu social media. Five models of machine learning have been proposed in this research which is NB, SVM, IBK, logistic and JRip. Other than that, four deep learning methods were also proposed for this task which is CNN, LSTM, bidirectional long short-term memory (BLSTM), and CLSTM. The result shows that CNN has a better result than other methods that have been used with an accuracy of 96.2% in Urdu and 91.4% in Roman Urdu.

Another research implemented a combined architecture of deep learning, i.e.: recurrent neural network (RNN) and LSTM for faults classification. The research succeeds to classify multi-label faults quite well even without preprocessing [16]. Du *et al.* also prove that the hybrid architecture of RNN and LSTM is able to classify claritin-october-twitter dataset with an accuracy that reached approximate 97% [17]. These studies' results prove that using a hybrid deep learning method to handle classification tasks presents better accuracy than a single deep learning approach. From the problem about the abusive comments issue on social media that was mentioned above, we need to identify abusive comments automatically by adopting the model from previous research to produce a good performance in the classification model.

This study aims to identify abusive comments by utilizing the hybrid deep learning approach for Indonesian social media data. The architecture used is a combination between RNN and LSTM. Both are complementary, the RNN algorithm is used to map the input sequence to a fixed-size vector to the hidden vector component which is used to summarize all the information in the previous process, then the LSTM algorithm is implemented to help overcome the gradient vector growth components that have the potential to exist in the RNN algorithm [18]. So that it can increase the performance of the abusive comment identification model.

2. STUDY OF LITERATURE

2.1. Sociolinguistic study

Abusive speech is an expression that is spoken either orally or in text and contains words or phrases that are rude or dirty, either in the context of jokes, a conversation of vulgar sex or of cursing someone [8]. Referring to sociolinguistic studies [19], Indonesian abusive word can be affected from daily conversation such as: i) describe an unpleasant situation or condition, e.g.: “*gila*” (in English: crazy), “*bodoh*” (in English: stupid), “*najis*” (in English: excrement), and “*celaka*” (in English: accurst); ii) compare animals characteristics with individual, e.g.: “*anjing*” (in English: dog), and “*babi*” (in English: pig); iii) abusive word that connect about astral being, e.g.: “*setan*” (in English: satan), and “*iblis*” (in English: devil); iv) depends on bad reference of that object, e.g.: “*tai*” (in English: shit), and “*gombel*” (in English: crap); v) body part that usually related with sex activity or another body part, e.g. “*matamu*” (*in English: your eyes*) cause someone make a mistake with their eyes; vi) express of displeasure or annoyed that related with family member usually add with suffix *-mu*, e.g.: “*bapakmu*” (in English: your father), “*kakekmu*” (in English: your grandpa), “*mbahmu*”) (in English: your grandma); and vii) related with profession using phrase about low profession and forbidden by religion, e.g.: “*maling*” (in English: thief), “*babu*” (in English: maid), “*lonte*” (in English: bitch).

2.2. Related works

Prabowo *et al* [20] proposed a classification process to recognize Indonesian abusive comments and hate speech on Twitter by implementing SVM. From the result, it was discovered that SVM with the help of the word unigram feature yields quite good results compared to other methods. However, the accuracy of the resulting system is still low around 68.43%.

Other researches were conducted to detect an abusive comment in Indonesian tweets using various machine learning techniques such as NB [8], binary relevance [13], and logistic regression [21]. Nevertheless, the result from those studies is not optimal enough to recognize hate speech from Twitter comments in Indonesia language with an average of accuracy under 80%. Table 1 shows existing research on Indonesian abusive comment classification using various machine learning techniques.

Table 1. Existing research on Indonesian abusive comment classification

Writers (Year)	Contribution	Method	Result
Ibrohim and Budi (2018) [8]	Classification of abusive in Indonesia language	Naïve Bayes	The accuracy of the resulting system is around 86.43%
Prabowo <i>et al.</i> (2019) [20]	Abusive comments and hate speech multi-classification on Twitter with Indonesia language	Support Vector Machine	However, the accuracy of the resulting system is not good enough around 68.43%.
Ibrohim and Budi (2019) [13]	Multi-classification of abusive and hate speech in Indonesia language	Binary Relevance	The accuracy of the system is around 73.53%
Ibrohim <i>et al.</i> (2019) [21]	Hate Speech and Abusive Language on Indonesian Twitter	Logistic Regression	The result of the system is 79.85%

To the best of our knowledge, research on Indonesian abusive comment classification using a deep learning approach is still limited and need to be explored. Neural network (NN) is part of machine learning that adapts hidden layer structures or implicit data patterns and is flexible for use in supervised, semi-supervised, or unsupervised learning [22]. NN has now been transformed into deep learning with excellent performance and can be implemented in various fields, including RNN and LSTM. Systematically, ANN is in the form of a graph with neurons or nodes (vertex) and synapses (edge) making it easier to explain operations on ANN in the form of linear algebraic notation. [22]. The concept of RNN is to create a network topology that can represent sequential or time-series data [22]. The main key of RNN is memorization [23]. LSTM is part of the RNN architecture which is specifically designed to model temporal sequences and their remote dependencies more accurately than conventional RNNs [24]. RNN functions to map input sequences to fixed-size vectors and LSTM is implemented to overcome gradient vector growth components that have the potential to exist in the RNN algorithm [18].

3. RESEARCH METHOD

This research used a classification task to identify abusive comments on Indonesian social media data. The deep learning approach is chosen with RNN and LSTM architectures referred to as Hybrid deep learning. The pipeline of the proposed model for abusive comments identification is shown in Figure 1.

The first step conducted is text preprocessing, then proceed with separating the dataset into train and test. Furthermore, the next step is modelling using the proposed hybrid deep learning architecture then validating the model before going to the implementation stage. Finally, the evaluation of the model is conducted to test the model using a confusion matrix table.

3.1. Dataset

The dataset that is used in this research is the Indonesian Abusive and Hate Speech from Kaggle with an amount of 13169 tweets. This research utilizes the target data in the Abusive column. The data is divided into two classes: abusive and non-abusive. The data distribution is as shown in Figure 2(a). There is an issue with the data related to imbalances. Therefore, the oversampling technique is necessary to overcome that problem. This is needed because the amount of data that is labelled before oversampling differs by almost 50%. This warrants for the oversampling technique to be conducted to increase abusive labelled data. Additionally, it is expected to improve system performance. The distribution of the data after the oversampling technique is carried out as shown in Figure 2(b).

3.2. Preprocessing

Preprocessing text involves case folding, filtering, stemming, tokenizing, sequencing and padding. The preprocessing phase started with lower case conversion and punctuation removal, specifically case

folding. The next step is stop word removing which is part of the filtering process. Afterwards, searching of the root word, namely stemming was conducted. In this research, we used Sastrawi stemmer. The last step is Tokenization which is the process of converting sentences into phrases, words or tokens that the system can understand [25].

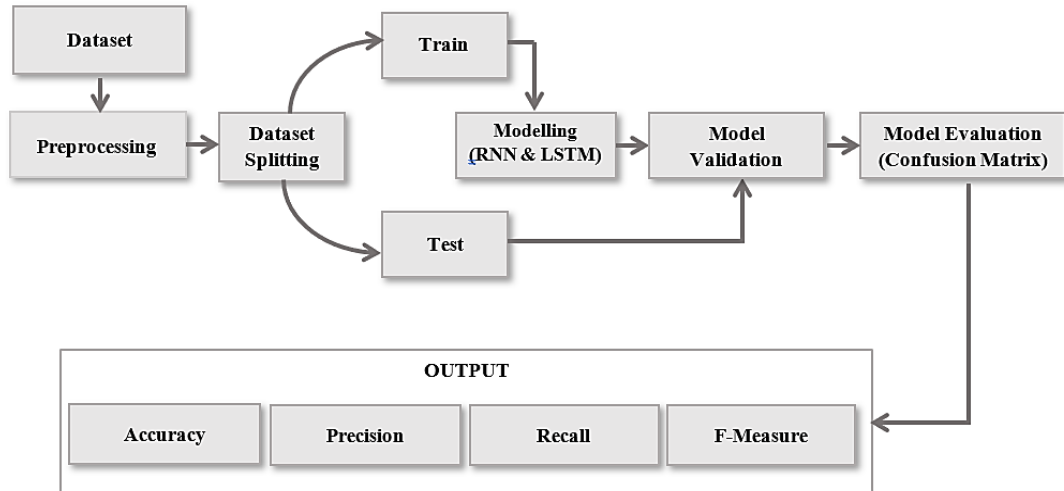


Figure 1. The pipeline of abusive comments identification using hybrid deep learning

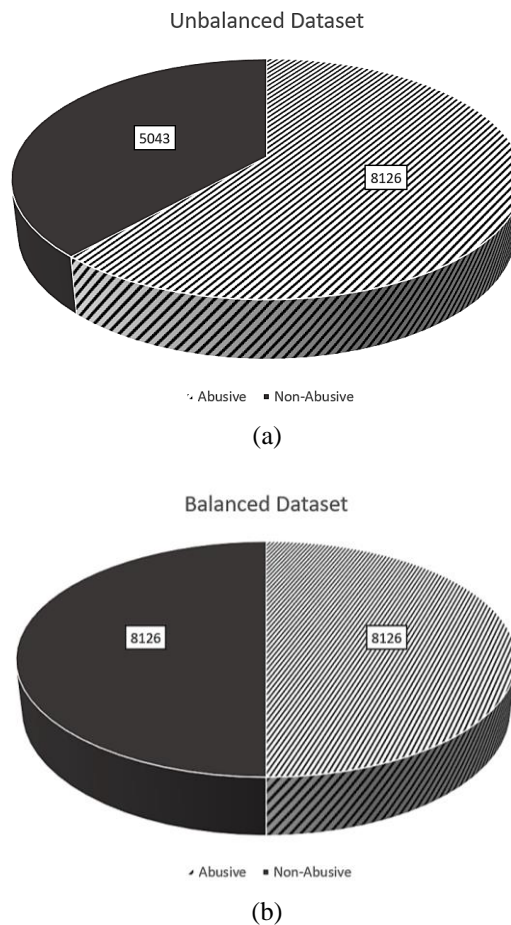


Figure 2. Data balancing visualization (a) before and (b) after oversampling

3.3. Data splitting

This process involves dividing the data into two sets for training and testing. The data is divided by the proportion of 80% data for the training set and 20% data for the testing set, with the total number of train data being 6501 for non-abusive and 6500 for abusive and the number of test data being 1626 for non-abusive and 1625 for abusive.

3.4. RNN-LSTM

In general, the RNN concept can be visualized in Figure 3 that accordance with the recurrent principle of remembering (memorizing) previous events. Meanwhile, LSTM is an architecture of RNN which was more accurate than the conventional RNN. LSTM is known to improve deficiencies found in conventional RNN. In the LSTM feature, there are several stages [26].

The detail of LSTM architecture is shown in Figure 4. Figure 4 depicts that LSTM has 3 gates, i.e. forgot gate, input gate, and output gate, which computation process for LSTM [27] is started with stored input value into the cell state when input gate allowed the process. In (1) and (2) show the calculation of the value input gate and possible value from the cell state.

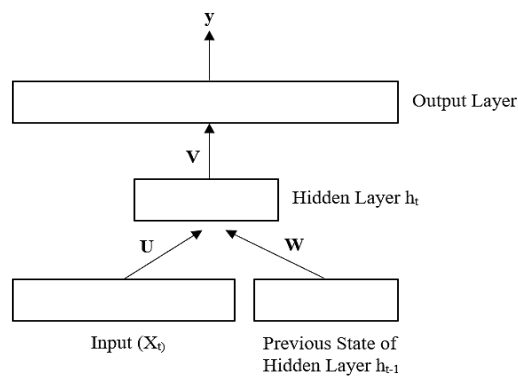


Figure 3. Concept of RNN

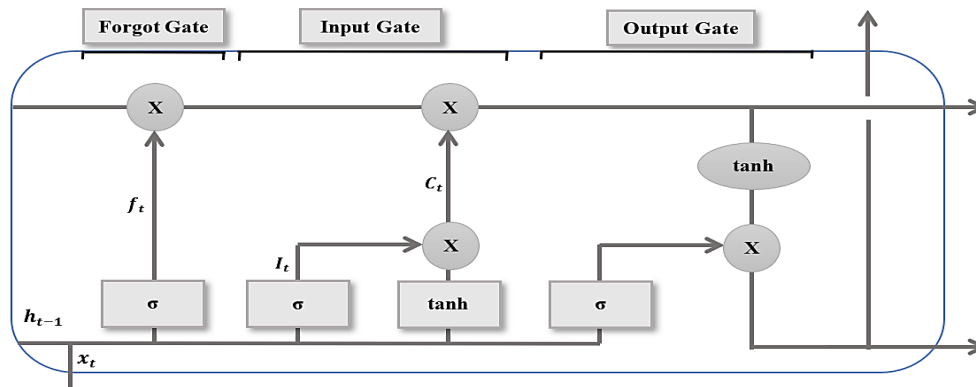


Figure 4. Architecture of LSTM

RNN concept can be described as a formula, as in (1). In that equation x_t is hidden state from input of time-t and h_{t-1} is hidden state in previous time. Whereas f is the activation function (non-linear, can be derived) [22]. The f function can be replaced with LSTM.

$$h_t = f(x_t, h_{t-1}, b) \tag{1}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{2}$$

i_t in (2) represent value input gate, W_i represent of weight value of i , x_t show value of the input in t condition, U_i represent of weight value of output at i . Then, h_{t-1} in (1) present value of output in $t - 1$ condition, b_i the present bias of input gate, and σ call with sigmoid function.

$$C_t = \tanh(W_c x_t + U_i h_{c-1} + b_c) \quad (3)$$

C_t the present possible value of *cell state*, W_c represent of weight value of the input in c condition, x_t show value of the input in t condition, U_i represent of weight value of output at i . Then, h_{c-1} in (3) present value of output in $c - 1$ condition, b_c the present bias of c , and \tanh can call as tangent function. Therefore, in (4) present process of forgot gate.

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (4)$$

Value of forget gate present with f_t , W_f represent of weight value of the input in f condition, x_t show value of the input in t condition, U_f represent of weight value of output at f , h_{t-1} in (4) present value of output in $t - 1$ condition, the bias of forget gate represent in b_f , and σ called as sigmoid function. Therefore, in (5) present process of cell state.

$$C_t = i_t \times C_t + f_t \times C_{t-1} \quad (5)$$

Value of cell state present with C_t , the value of input gate value shown at i_t , C_t present possible of cell state, the value of forget gate present in f_t and value of previous cell state show with C_{t-1} . The next process after generating new memory of cell state has been done, the process of output gate started in (6).

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (6)$$

The value that represents of output gate is o_t , W_o represent of weight value of the input in o condition, x_t show value of the input in t condition, U_o show value of the input in o condition, h_{t-1} in (6) present value of output in $t - 1$ condition, output gate bias present with b_o , and sigmoid function represents at σ symbol. The final output process shows in (7).

$$h_t = o_t * \tanh(C_t) \quad (7)$$

The final output represented with h_t and output gate represent in o_t value, C_t present of new memory cell state value and \tanh called as tangent function.

3.4. RNN-LSTM implementation

In the modelling process, the combination of RNN and LSTM is used to develop the optimal model. The train data is entered into the RNN and LSTM classification models as a data input that has been created in Figure 5. The input layer is defined as an embedding layer, where in this layer there is a vocabulary retrieval process that is coded with an array of integers and embedding vectors for each word index.

The resulting dimensions are batch size, sequence, and embedding size. Dimension of embedding sets the number of features for the word i.e., the number of hidden units. The embedding result is in the form of a matrix with dictionary length and embedding dimensions. Then, added a RNN Layer with *SimpleRNN* from tensorflow module. The next step is to enter a LSTM memory block, the LSTM contain a special unit called a memory block in the recurrent hidden layer that detail explained in Figures 3 and 4. Then, added layer dense to create a complexity NN layer and dropout to handle overfitting problems [28]. The last step is data output which is represented with a binary value (0 and 1). These values are the probability, 0 is for non-abusive data and 1 for abusive data. This label is used to classify abusive sentences.

3.5. Model validation

The model that has been built should be validated with data validation that is entered in *model.fit()* function with x variable is an input data which contain *train_sequence* while y variable is a target data. Both variables can be assigned with Numpy array or Tensorflow tensor while y variable filled with *train_labels*. The next parameter is *batch_size*, a number of samples per gradient update and that contain integer on none, we use 32 as a default of *batch_size*. For Epoch, a number of iterations in train model, we use 10 epochs to train the model. Meanwhile, Callback using *rlrp* is applied during training and last verbose using integer 1 that means progress bar in tensorflow documentation. And after that step, the system will be showing the

accuracy of the training model. Finally, the modelling is done and produced the values of the training process, e.i.: loss and accuracy of each epoch.

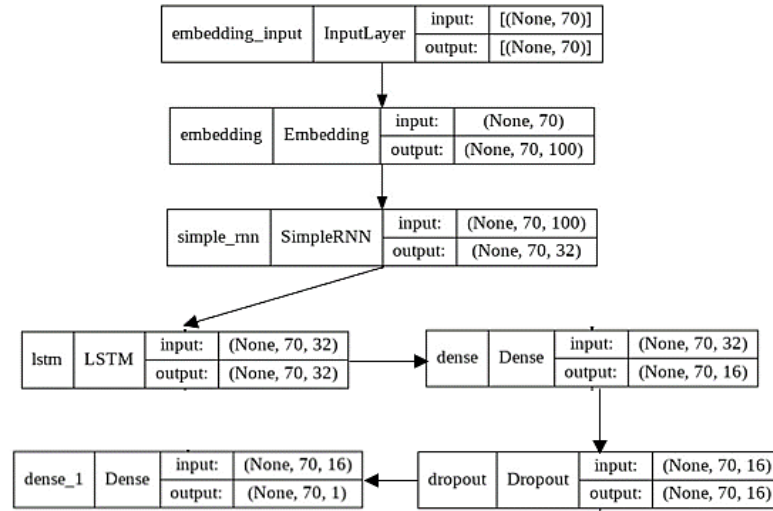


Figure 5. The architecture of RNN and LSTM

3.6. Model evaluation

The evaluation measurement is used to test the reliability of the model to get accuracy, precision, recall and f-measure values based on the confusion matrix table. The confusion matrix summarizes the classification performance of the system in actual and predicted form through the entered test data [29]. From the confusion matrix table, the accuracy, precision, recall and f-measure score can be calculated using (8)-(11). It means when the system predicts a word as an abusive class and actually declared as abusive then the statement is represented as true positive (TP). If the system predicts a word as a non-abusive class but actually in the dataset the sentence is stated rudely then the statement is represented with false negative (FN). Meanwhile, the system predicts a word as an abusive class but actually declared as a non-abusive then the statement is represented as false positive (FP). However, when the system predicts a word as a non-abusive class and actually declared as a non-abusive then the statement is represented as true negative (TN).

$$\text{Accuracy} = \frac{TP+TN}{(TP+FP+FN+TN)} \quad (8)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (9)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (10)$$

$$f \text{ measure} = \frac{2 \times (\text{Recall} \times \text{Precision})}{\text{Recall} + \text{Precision}} \quad (11)$$

4. RESULTS AND DISCUSSION

This section shows the results obtained during the training, validation, testing process and presents the important findings obtained from these results. One of which is the effect of using the RNN-LSTM hybrid deep learning model and the impact of balanced data. The distribution of the previous data shows that the data used is unbalanced by using the random oversampling technique. This technique takes some random data from the abuse category with an adjusted amount based on the largest number of classes that will be used as data into balanced proportion as in Table 2. The data after oversampling and the source code about the research that distributed on *github* [30].

The use of oversampling can provide a fairly good increase in system performance. The use of balanced data greatly affects the system in classifying a text, the results of the comparison between the use of oversampling and data that did not go through the oversampling process. The results obtained provide an

increase of 6% to 8% for precision, recall, and f1-score as shown in Table 3. Meanwhile, The use of RNN-LSTM also provides an increase in the use of several machine learning methods in previous studies [8]. As in Table 4 which shows the system performance using f1-score, the previous method using NB shows the highest average results of 86% and the use of RNN-LSTM gives an increase of 8% for the f1-score value.

The use of RNN-LSTM with parameter settings in Figure 5 which has a fairly complex layer arrangement was able to give good results in classifying coarse text. However, the layer is too complex causes overfitting in the model [28] as shown in the graph in Figure 6. Overfitting is a condition when a model learns the training dataset too well but does not perform well on a data test [31].

Furthermore, through the results of the confusion matrix, the ability of the model to predict the test data given can be known. The proposed model can provide correct predictions of 1521 of 1626 test data given in the non-abusive category. As for the rough category, the system can provide correct predictions as many as 1537 of the 1625 test data provided. There are quite a lot of correct data predicted by the system, this is reinforced by classifying it with test data that comes from outside the dataset used, as in Table 4. Then, Table 5 shows a row of sentences along with the probability values issued by the system based on the model that has been built and trained previously. The higher the probability value, it indicates that the sentence contains abusive words quite high. Meanwhile, the smaller the probability, the lower the abusive word in a sentence.

Table 2. Proportion of Data

Before Oversampling		After Oversampling	
Abusive	Non-Abusive	Abusive	Non-Abusive
5043	8126	8126	8126

Table 3. Comparison of oversampling and no-oversampling

System Performance	No-Oversampling	Oversampling
Precision	88%	94%
Recall	86%	95%
F1-Score	87%	94%

Table 4. Comparison of proposed method with previous research

Method	F1-Score
NB with word unigram [8]	86.43%
NB with unigram+bigram [8]	86.12%
NB with trigram+quadgram [8]	86.17%
RNN-LSTM with oversampling (proposed)	94.00%

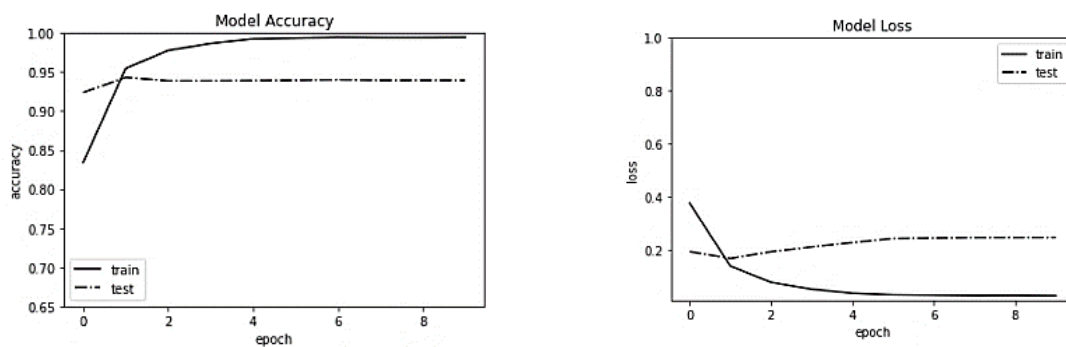


Figure 6. Graph accuracy and loss for RNN-LSTM model

Table 5. Test Model with Other Data Test

Sentences	Probability
<i>najis ih</i> (dirty)	93.38%
<i>wkwkwk jijik euh</i> (mucky)	04.92%
<i>dia memang kurang berpendidikan</i> (less educated)	86.71%
<i>otaknya ngga dipakai</i> (the brain is not used)	13.62%

5. CONCLUSION

The results of the study are used to classify rude comments on social media, especially twitter using a hybrid RNN and LSTM architecture. From the test results, the hybrid architecture RNN and LSTM can form a system that can identify abusive comments in comments uploaded on Twitter with optimal performance as evidenced by the results of precision, recall and f-measure with each number is 94%, 95% and 94% respectively. The confusion matrix also displays the system's performance in predicting the test data provided is quite good, 1537 abusive data can be predicted correctly by the system from a total of 1652 test data provided. Moreover, when the system test on other datasets, the system can predict better by showing a probability of abusive words. In addition, the use of oversampling techniques to handle imbalanced data can also contribute to improving system performance by 4% and a significant increase in precision, recall and f-measure by 6%, 9%, and 7% respectively for abusive data. The proposed model can carry out initial identification of cyberbullying through the classification of abusive comments on social media. For future work, it can be further developed to build an automatic blocking system in support of government programs regarding cyberbullying prevention. In future work, our research will focus on trying other combinations of deep learning methods, such as LSTM-CNN, BiLSTM, and several methods that are considered more sophisticated to improve system performance, especially in reducing overfit states in the model. In addition, we will try to implement this method on a larger amount of data to see if the method is able to produce good performance on data that has much more capacity.




REFERENCES

- [1] M. Amin *et al.*, "Security and privacy awareness of smartphone users in Indonesia," *J. Phys. Conf. Ser.*, vol. 1882, no. 1, p. 12134, May 2021, doi: 10.1088/1742-6596/1882/1/012134.
- [2] S. D. A. Putri, M. O. Ibrohim, and I. Budi, "Abusive language and hate speech detection for indonesian-local language in social media text," in *Recent Advances in Information and Communication Technology 2021*, 2021, pp. 88–98.
- [3] N. Cécillon, V. Labatut, R. Dufour, and G. Linarès, "Graph embeddings for abusive language detection," *SN Comput. Sci.*, vol. 2, no. 1, p. 37, Feb. 2021, doi: 10.1007/s42979-020-00413-7.
- [4] H. Gong, A. Valido, K. M. Ingram, G. Fanti, S. Bhat, and D. L. Espelage, "Abusive language detection in heterogeneous contexts: dataset collection and the role of supervised attention," 2021, [Online]. Available: <http://arxiv.org/abs/2105.11119>.
- [5] S. E. Saad and J. Yang, "Twitter sentiment analysis based on ordinal regression," *IEEE Access*, vol. 7, pp. 163677–163685, 2019, doi: 10.1109/ACCESS.2019.2952127.
- [6] S. D. Swamy, A. Jamatia, and B. Gambäck, "Studying generalisability across abusive language detection datasets," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 2019, pp. 940–950, doi: 10.18653/v1/K19-1088.
- [7] S. Tuarob and J. L. Mitranont, "Automatic discovery of abusive thai language usages in social networks," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, pp. 267–278.
- [8] M. O. Ibrohim and I. Budi, "A dataset and preliminaries study for abusive language detection in Indonesian social media," *Procedia Comput. Sci.*, vol. 135, pp. 222–229, 2018, doi: 10.1016/j.procs.2018.08.169.
- [9] I. Krisnana *et al.*, "Adolescent characteristics and parenting style as the determinant factors of bullying in Indonesia: a cross-sectional study," *Int. J. Adolesc. Med. Health*, vol. 33, no. 5, p. 1, Oct. 2021, doi: 10.1515/ijamh-2019-0019.
- [10] R. Wahanisa, R. Prihastuty, and M. Dzikirullah H. Noho, "Preventive measures of cyberbullying on adolescents in indonesia: a legal analysis," *Lentera Huk.*, vol. 8, no. 2, p. 267, Jul. 2021, doi: 10.19184/ejlh.v8i2.23503.
- [11] J. H. Park and P. Fung, "One-step and two-step classification for abusive language detection on Twitter," in *Proceedings of the First Workshop on Abusive Language Online*, 2017, pp. 41–45, doi: 10.18653/v1/W17-3006.
- [12] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings of the 25th International Conference on World Wide Web*, Apr. 2016, pp. 145–153, doi: 10.1145/2872427.2883062.
- [13] M. O. Ibrohim and I. Budi, "Multi-label hate speech and abusive language detection in Indonesian twitter," in *Proceedings of the Third Workshop on Abusive Language Online*, 2019, pp. 46–57, doi: 10.18653/v1/W19-3506.
- [14] P. Chakraborty and M. H. Seddiqui, "Threat and abusive language detection on social media in Bengali language," in *1st International Conference on Advances in Science, Engineering and Robotics Technology 2019, ICASERT 2019*, 2019, pp. 1–6, doi: 10.1109/ICASERT.2019.8934609.
- [15] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. AbdelMajeed, and T. Zia, "Abusive language detection from social media comments using conventional machine learning and deep learning approaches," *Multimed. Syst.*, no. 0123456789, Apr. 2021, doi: 10.1007/s00530-021-00784-8.
- [16] G. S. Chadha, A. Panambilly, A. Schwung, and S. X. Ding, "Bidirectional deep recurrent neural networks for process fault classification," *ISA Trans.*, vol. 106, pp. 330–342, Nov. 2020, doi: 10.1016/j.isatra.2020.07.011.
- [17] J. Du, C.-M. Vong, and C. L. P. Chen, "Novel efficient RNN and LSTM-like architectures: recurrent and gated broad learning systems and their applications for text classification," *IEEE Trans. Cybern.*, vol. 51, no. 3, pp. 1586–1597, Mar. 2021, doi: 10.1109/TCYB.2020.2969705.
- [18] L. Kurniasari and A. Setyanto, "Sentiment analysis using recurrent neural network-lstm in bahasa Indonesia," *J. Eng. Sci. Technol.*, vol. 15, no. 5, pp. 3242–3256, 2020.
- [19] F. Priyanto and A. Ashadi, "The acquisition of swear words by students in Central Kalimantan," *RETORIKA J. Bahasa, Sastra, dan Pengajarannya*, vol. 13, no. 2, pp. 1–26, Aug. 2020, doi: 10.26858/retorika.v13i2.13803.
- [20] F. A. Prabowo, M. O. Ibrohim, and I. Budi, "Hierarchical multi-label classification to identify hate speech and abusive language on Indonesian Twitter," in *2019 6th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)*, Sep. 2019, pp. 1–5, doi: 10.1109/ICITACEE.2019.8904425.
- [21] M. O. Ibrohim, M. A. Setiadi, and I. Budi, "Identification of hate speech and abusive language on indonesian Twitter using the Word2vec, part of speech and emoji features," in *Proceedings of the International Conference on Advanced Information Science and System*, Nov. 2019, pp. 1–5, doi: 10.1145/3373477.3373495.




- [22] Z. Ahmad, A. S. Khan, C. W. Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," *Trans. Emerg. Telecommun. Technol.*, vol. 32, no. 1, pp. 1–235, Jan. 2021, doi: 10.1002/ett.4150.
- [23] S. Marsella and J. Gratch, "Computationally modeling human emotion," *Commun. ACM*, vol. 57, no. 12, pp. 56–67, Nov. 2014, doi: 10.1145/2631912.
- [24] D. Lee *et al.*, "Long short-term memory recurrent neural network-based acoustic model using connectionist temporal classification on a large-scale training corpus," *China Commun.*, vol. 14, no. 9, pp. 23–31, Sep. 2017, doi: 10.1109/CC.2017.8068761.
- [25] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," *Inf. Process. Manag.*, vol. 50, no. 1, pp. 104–112, Jan. 2014, doi: 10.1016/j.ipm.2013.08.006.
- [26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
- [27] H. Chung and K. Shin, "Genetic algorithm-optimized long short-term memory network for stock market prediction," *Sustainability*, vol. 10, no. 10, p. 3765, Oct. 2018, doi: 10.3390/su10103765.
- [28] H. il Lim, "A study on dropout techniques to reduce overfitting in deep neural networks," in *Lecture Notes in Electrical Engineering*, 2021, vol. 716, pp. 133–139, doi: 10.1007/978-981-15-9309-3_20.
- [29] D. Chicco, V. Starovoitov, and G. Jurman, "The benefits of the matthews correlation coefficient (MCC) over the diagnostic odds ratio (DOR) in binary classification assessment," *IEEE Access*, vol. 9, no. Mcc, pp. 47112–47124, 2021, doi: 10.1109/ACCESS.2021.3068614.
- [30] T. I. Sari, "Abusive Twitter identification fithub." [Online]. Available: <https://github.com/tiaraintana/Abusive-Twitter-Identification>.
- [31] A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra, "Grokking: generalization beyond overfitting on small algorithmic datasets," pp. 1–10, Jan. 2022, [Online]. Available: <http://arxiv.org/abs/2201.02177>.

BIOGRAPHIES OF AUTHORS






Tiara Intana Sari    is a student who are pursuing undergraduate studies in the Department of Informatics, Faculty of Engineering at the University of Muhammadiyah Malang. Her area of interest is Data Science. She can be contacted at email: tiaraintana@webmail.umm.ac.id.






Zalfa Natania Ardilla    is a student of Bachelor degree at Informatics Department of Engineering Faculty, University of Muhammadiyah Malang. Her area of interest in data science. She can be contacted at email: zalfaardilla@webmail.umm.ac.id.



Nur Hayatin    is a lecturer at the Informatics Department of Engineering Faculty, University of Muhammadiyah Malang, Indonesia. She received her Master in Informatics Engineering from the Institute of Technology Sepuluh Nopember Surabaya Indonesia. Currently, she is undertaking a Graduate Research Assistant program at Universiti Malaysia Sabah. Her area of interest is data science specific in Natural Language Processing, social media analytics, Data Mining, and Information Retrieval. She can be contacted at email: noorhayatin@umm.ac.id.



Ruhaila Maskat    is a senior lecturer at the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Shah Alam, Malaysia. In 2016, she was awarded a Ph.D. in Computer Science from the University of Manchester, United Kingdom. Her research interest then was in Pay-As-You-Go dataspace which later evolved to Data Science where she is now an EMC Dell Data Associate as well as holding four other professional certifications from RapidMiner in the areas of machine learning and data engineering. Recently, she was awarded the Kaggle BIPOC grant. Her current research grant with the Malaysian government involves conducting analytics on social media text to detect mental illness. She can be contacted at email: ruhaila@fskm.uitm.edu.my.