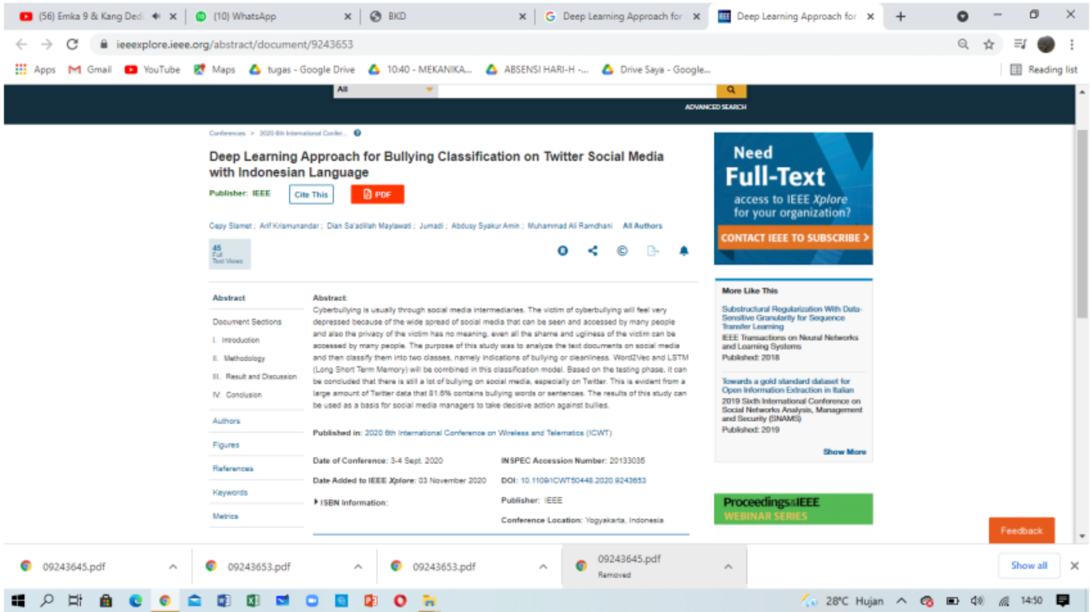
Deep Learning Approach for Bullying Classification on Twitter Social Media with Indonesian Language

by Abdusy Syakur Amin -



ieeexplore.ieee.org/abstract/document/9243653

Conferences > 2020 6th International Confer...

Deep Learning Approach for Bullying Classification on Twitter Social Media with Indonesian Language

Publisher: IEEE | Cite This | PDF

Cepy Slamet ; Arif Krismunandar ; Dian Sa'adillah Maylawati ; Jumadi ; Abdusy Syakur Amin ; Muhammad Ali Ramdhani | All Authors

45 Full Text Views

Abstract

Document Sections

I. Introduction

II. Methodology

III. Result and Discussion

IV. Conclusion

Authors

Figures

References

Keywords

Metrics

Abstract:
Cyberbullying is usually through social media intermediaries. The victim of cyberbullying will feel very depressed because of the wide spread of social media that can be seen and accessed by many people and also the privacy of the victim has no meaning, even all the shame and ugliness of the victim can be accessed by many people. The purpose of this study was to analyze the text documents on social media and then classify them into two classes, namely indications of bullying or cleanliness. Word2Vec and LSTM (Long Short Term Memory) will be combined in this classification model. Based on the testing phase, it can be concluded that there is still a lot of bullying on social media, especially on Twitter. This is evident from a large amount of Twitter data that 81.6% contains bullying words or sentences. The results of this study can be used as a basis for social media managers to take decisive action against bullies.

Published in: 2020 6th International Conference on Wireless and Telematics (ICWT)

Date of Conference: 3-4 Sept. 2020

Date Added to IEEE Xplore: 03 November 2020

ISBN Information:

INSPEC Accession Number: 20133036

DOI: 10.1109/ICWT50448.2020.9243653

Publisher: IEEE

Conference Location: Yogyakarta, Indonesia

Need Full-Text access to IEEE Xplore for your organization? CONTACT IEEE TO SUBSCRIBE >

More Like This

Substructural Regularization With Data-Sensitive Granularity for Sequence Transfer Learning
IEEE Transactions on Neural Networks and Learning Systems
Published: 2018

Towards a gold standard dataset for Open Information Extraction in Italian
2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)
Published: 2019

Show More

Proceedings of the IEEE WEBINAR SERIES

Feedback

Deep Learning Approach for Bullying Classification on Twitter Social Media with Indonesian Language

Cepy Slamet
Department of Informatics
UIN Sunan Gunung Djati Bandung
Bandung, Indonesia
cepy_lucky@uinsgd.ac.id

Arif Krismunandar
Department of Informatics
UIN Sunan Gunung Djati Bandung
Bandung, Indonesia
arifkrismunandar@gmail.com

Dian Sa'adillah Maylawati
Center for Advanced Computing Technology
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia
p031810030@student.utem.edu.my

Jumadi
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
jumadi@uinsgd.ac.id

Abdusy Syakur Amin
Department of Industial Engineering
Universitas Pasundan
Bandung, Indonesia
abdusy.syakur@unpas.ac.id

Muhammad Ali Ramdhani
Department of Informatics
UIN Sunan Gunung Djati Bandung
Bandung, Indonesia
m_ali_ramdhani@uinsgd.ac.id

***Abstract*—Cyberbullying is usually through social media intermediaries. The victim of cyberbullying will feel very depressed because of the wide spread of social media that can be seen and accessed by many people and also the privacy of the victim has no meaning, even all the shame and ugliness of the victim can be accessed by many people. The purpose of this study was to analyze the text documents on social media and then classify them into two classes, namely indications of bullying or cleanliness. Word2Vec and LSTM (Long Short Term Memory) will be combined in this classification model. Based on the testing phase, it can be concluded that there is still a lot of bullying on social media, especially on Twitter. This is evident from a large amount of Twitter data that 81.6% contains bullying words or sentences. The results of this study can be used as a basis for social media managers to take decisive action against bullies.**

***Keywords*— deep learning, bullying, classification, text mining, LSTM, Long Short Term Memory, neural network, Word2Vec**

I. Introduction

Bullying is one of the detrimental actions that can interfere with the comfort and hurt a person, both physically and mentally, with the difference in strength and psychology on the part of the victim and the perpetrator which is done repeatedly. Bullying can be divided into two media, namely traditional bullying and cyberbullying. Traditional bullying occurs because of direct physical contact or verbal speech between the victim and the perpetrator, for example, beatings, expressing swearing or harsh sentences to the victim, and others, worse traditional bullying is usually done by a group of people who bully someone because of their physical imperfections. or the nature of the victim to be either mentally retarded or otherwise. Cyberbullying occurs because of the existence of social media intermediaries and victims who are harassed and persecuted through social media, usually cyberbullying victims will feel very depressed due to the widespread of social media that can be seen and accessed by many people and also the privacy of the victim has no meaning even all disgrace and the ugliness of the victim is accessible to many people. The worst impact for victims who cannot withstand the burden of bullying is that the victim can take a reckless decision by ending his life in other words, suicide [1]–[6].

Based on previous research that has been there, victims of bullying are usually people who have problems in terms of health (physical), emotional (psychological), and they work together in groups. Usually, these victims (bullying) are reported to have high levels of anxiety, are prone to depression, and also usually have very low self-esteem because they are usually under stress and put them on the bad side. Based on other research, the level of bullying in this traditional method of bullying is in the range of 9.68% - 89.6% with a range of bullying victims ranging from around 9% - 97.9%. Whereas in the cyberbullying method, the level of bullying reached the range of 5.3% - 31.5% with the number of bullying victims around 2.2% - 56.2% [3].

Based on the above research, there is also research which states that the bullying case in the test involved 7.508 participants who had experienced being victims of bullying, both traditional bullying and also cyberbullying, and 326 participants or about 4.3% of cases of which were caused by demographic factors, parental factors. (orphaned) and also the factors that belong to the victim. The study was conducted on participants aged 14.3 years [3].

Cyberbullying is also the fault of every information technology user who is not wise by harming or hurting and harassing other individuals repeatedly without thinking about the impact and consequences of what he has done to others, both the psychological trauma he received. Cyberbullying can involve a group that knows each other and a group that doesn't know each other. Cyberbullying can make the bully use a false identity which can provide a sense of freedom from legal bondage because he feels his identity is disguised without thinking about the existence of social rules and moral values that should be preserved in the wise use of social media.

978-1-7281-7596-6/20/$31.00 ©2020 IEEE

Cyberbullying usually occurs on social media such as Twitter, Instagram, Facebook, and others [1], [2], [7].

In Indonesia, the misuse of social media has been regulated in the Electronic Information Law or which can be abbreviated (UU ITE) number 11 of 2008 article 27 paragraph 3 which describes the spread of good names or insults. Cyberbullying can be classified into this category where there is an element of humiliation in it. At this time, KOMINFO is collaborating with Google and Twitter in reducing or eliminating content that contains sensitive elements and harms someone in the form of hoax news, pornography, cyberbullying, and others. KOMINFO hopes that in the future the spread of such negative content can be resolved quickly. On the other hand, Twitter also provides a function where we can report content containing negative content separately so that it can be processed more quickly. Therefore, with the availability of an automatic system for identifying cyberbullying tweets, it is hoped that this can help speed up the follow-up of tweets that contain negative things more efficiently and quickly.

The purpose of this research is to take a Natural Language Processing (NLP) approach to recognize and extract text documents to be classified as bullying or not. One of NLP technique is Sentiment Analysis which analyze the data based on positive, negative, or neutral opinion [8], [9]. Long Short Term Memory Neural Network (LSTM) is one type of Recurrent Neural Network (RNN) [10]–[14]. LSTM stores information in the form of existing patterns in data [15]–[17]. The ability of LSTM is that it can learn what data will be stored and what data will be discarded, because, in each neuron, LSTM has several gates which have control to regulate memory in each part of the neuron itself. LSTMs are frequently used in processing video, text, and time-series data and has implemented on any case [18]–[21] especially text document [10]. Word2vec is a technique for NLP. It uses a neural network model to learn word associations from a large corpus of text. This algorithm will implement on processing phase. On previous work, it mainly use on sentiment analysis case [22]–[28] and on several works it combine with deep learning [29].

II. Methodology

A. Data Understanding

The data used is tweet data that contains sentences or elements of bullying with the keywords dog, bastard, and idiot. Tweet data collection was carried out from January 1, 2020, to April 3, 2020, the categorization of this data had previously been consulted with experts in the field of psychology who could assist in classifying this data themselves.

Before this data is processed, it must be processed first with a preprocessing process (Figure 1) which results in a clean and noise-free dataset so that it can be used for word weighting or converting words into vectors, this processing is carried out because it will test the accuracy of the Long Short Term Memory algorithm (LSTM). This algorithm will calculate the vectors it receives which is then separated by the hyper lane to produce a classification.

The dataset that has gone through the embedding process will then be processed by the system and then a separate training and testing process is carried out which results in the accuracy of the classification process carried out by the LSTM algorithm.

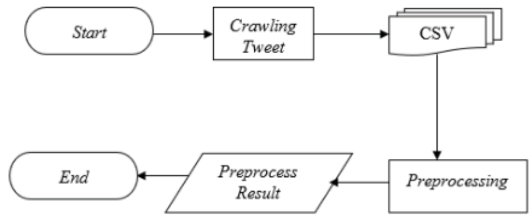


Fig. 1. Data Preparation Phase

Training data is data that will be used as training data for the system which will later produce a model. While the test data is data that will be used as testing data on the system, which will determine the level of accuracy of the classification process using the Long Short Term Memory (LSTM) algorithm. The data that has been categorized is then divided into 2 parts, namely training data and test data on each dataset, as we can see in the Figure 2.

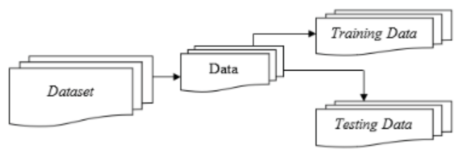


Fig. 2. Data Training and Testing

B. Text Preprocessing

The data that has been divided into two parts, namely test data and training data, then proceed to the preprocessing stage. Preprocessing is the stage of processing unstructured data into structured data so that the data can be more efficient for the processing stage. In this preprocessing, there are several steps to produce perfect data, namely case-folding, cleaning, stemming, stopword removal, and tokenizing.

C. Tokenizing

One of the word vector representations created by google which is named Word2Vec. With Word2Vec we can measure vectors for comparison. If we measure the distance between the vectors for the words "Indonesia" and "Jakarta" and the vectors for the words "Australia" and "Sydney", we will find that the distances will appear with a corresponding number. This is because the two words above are the name of a country and its capital. Because the meaning of the two sentences depends, the vector value of the word above is connected.

978-1-7281-7596-6/20/$31.00 ©2020 IEEE

D. Long Short Term Memory (LSTM) Algorithm

The first step used in the LSTM algorithm process is to determine which information will be removed or deleted from the electronic state. This decision is determined by the sigmoid layer "forget gate layer". The forget gate layer will process h_{t-1} and x_t as input in the form of numbers 1 or 0 in cell state c_{t-1}.

The next step is to determine what information will be stored in the cell state. This step is divided into two parts. In the first part, the sigmoid layer called the input gate layer will determine which values need to be updated. Next, the layer will create a candidate with a new value, $\tilde{C}_t$, which can be added to the cell state. The second stage is the output or the result of the input gate layer and the layer will be combined to update the cell state.

Next is to update the old state cell, C_{t-1}, to the new state cell, C_t. By multiplying by ft, in removing the information previously specified in the forget gate layer process. Then added with $i_t * \tilde{C}_t$, which becomes a new value and is used to update the state.

The last step in the LSTM method is that the output that comes out must match the cell state that has been processed first. First, the sigmoid layer will decide which part of the cell state will be output. Next, the output from the cell state is entered into the layer and multiplied by the sigmoid gate, with the aim that the resulting output is in accordance with what we have previously decided.

The image below shows how the process occurs in the LSTM algorithm. The next step after the Word2Vec process is weighting each existing word which then results from the Word2Vec process into a data format that can be processed and understood by the LSTM algorithm, which is then divided into 2 as training data and test data whose results can be seen after experimenting with the LSTM algorithm. Figure 3 illustrates the LSTM model.

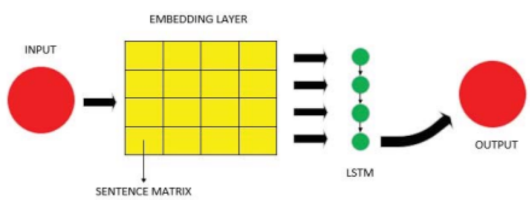


Fig. 3. LSTM Model

III. Result and Discussion

We conducted 5 test scenarios by dividing training data and testing data and dividing them in scenario 1, namely 90% training data and 10% testing data, scenario 2 which is 80% training data and 20% testing data, scenario 3 which is 70% training data and 30 % testing data, then 60% and 40%, and finally 50% testing data and 50% training data. There are 1359 datasets used for research on cyberbullying on social media twitter with Indonesian language. The Figure 4 presents the percentage of the data used in this analysis process. Where 0 is negative (means contain the bullying), while 1 is positive.

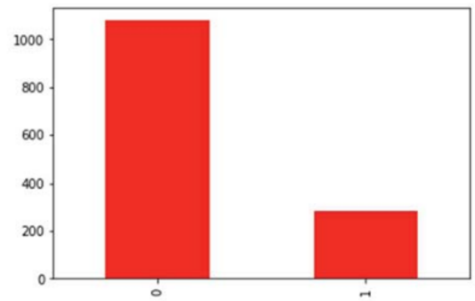


Fig. 4. Percentage of training data that contain bullying and do not contain bullying

A. Scenario 1 Testing

The first test carried out the amount of training data as much as 90% and testing data by 10%. The data being tested will display the first test.

$$Precission\ (positive) = \frac{92}{105} = 0{,}87 \tag{1}$$

$$Precission\ (ngative) = \frac{22}{31} = 0{,}70 \tag{2}$$

Then calculate the recall for each predicted class. Recall or sensitivity is the proportion of real positive cases predicted to be correct. The results of the recall calculation can be seen and substituted in the calculation below:

$$Recall\ (+) = \frac{92}{101} = 0{,}91 \tag{3}$$

$$Recall\ (-) = \frac{22}{35} = 0{,}62 \tag{4}$$

Then calculate the f1-score value as follows:

$$f1\ score\ (+) = 2\,x\,\frac{0{,}87\ x\ 0.95}{0{,}87 +\ 0.95} = 0.88 \tag{5}$$

$$f1\ score\ (-) = 2\,x\,\frac{0{,}70\ x\ 0.62}{0{,}70 +\ 0.62} = 0.65 \tag{6}$$

The last is calculating the accuracy of the sentiment classifying application. Accuracy can be determined by dividing the number of correct classifications with all documents classified.

$$Accuracy = \frac{92 + 22}{92 +\ 22 + 18 + 19}\ x\ 100\% = 83\% \tag{7}$$

ROC score of scenario 1 testing described in the Figure 5, while Confusion matrix of scenario 1 described in the Figure 6.

978-1-7281-7596-6/20/$31.00 ©2020 IEEE

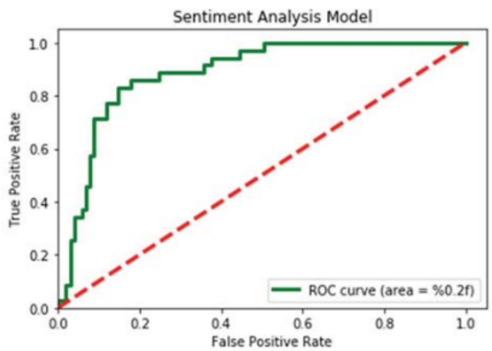


Fig. 5. The example of ROC Score of Scenario 1

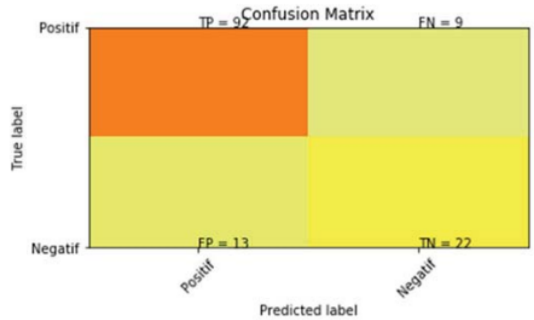


Fig. 6. The example of Confusion Matrix Score for Scenario 1

TABLE I. SUMMARY OF TESTING

Training Data		***Testing Data***		***Precision***		***Recall***		***F1-Score***		***Accuracy***
%	***Data***	***%***	***Data***	***0***	***1***	***0***	***1***	***0***	***1***	***%***
90%	1223	10%	136	70%	87%	62%	91%	65%	88%	83
80%	1087	20%	272	80%	87%	56%	95%	64%	90%	86
70%	951	30%	408	63%	82%	28%	95%	37%	87%	80
60%	815	40%	544	54%	80%	11%	97%	15%	87%	79
50%	679	50%	680	56%	98%	21%	95%	28%	96%	80
Mean of Accuracy				**64.6%**	**86.8%**	**35.6%**	**94.6%**	**41.8%**	**89.6%**	**81.6%**

B. Testing Summary

Tests that have been carried out in a total of 5 scenarios by producing the best accuracy value with time are efficient, therefore the test results are analyzed. Summary of testing describe on Table 1.

The graph in the Figure 7 shows that the first test was at 83%, then the second test was carried out, the accuracy increased to 86%. Then the third test has decreased to 80% then fell back 79%. Then the last test got an accuracy of 80%.

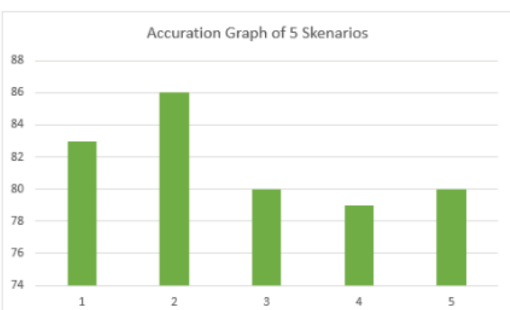


Fig. 7. Summary Graph of 5 Testing Scenarios

Thus, it is clear that the predictions generated from the Long Short Term Memory algorithm depend on the amount of data used as training data and test data. This is because the average value in the first experiment to the third experiment decreases with the same number of epochs. Then in the fourth and fifth experiments, the epoch is added to train the more data. After that, the accuracy increases even though the training data used is less than the previous training data.

IV. CONCLUSION

Based on the results of this study, it can be concluded that there is still a lot of bullying on social media, especially on Twitter. This is evident from a large amount of Twitter data that 81.6% contains bullying words or sentences. The results of this study can be used as a basis for social media managers to take decisive action against bullies, especially on social media.

Factors that can increase this accuracy value are the first data used, the selection of randomly randomized data on training data and testing data, noise or the cleanliness of the data itself, factors that can be used to support this research are accurate labeling (checking repeatedly). Another factor that can affect accuracy is the number of documents used. The more documents used, the better the accuracy value. Models will learn more and more and produce high accuracy, noise-free data, and repeated labeling to ensure that the classified data is correct and there are no errors in the data.

Further works, we suggest adding more documents. This can increase the value for better accuracy. It is advisable to carry out supervision by a linguist and check manually the data set that will be used in the training and testing process so that the documents to be processed are completely clean from noise and other errors.

978-1-7281-7596-6/20/$31.00 ©2020 IEEE

REFERENCES

[1] Q. Li, Y. Luo, Z. Hao, B. Smith, Y. Guo, and T. Cheng, "Author Correction: Risk Factors of Cyberbullying Perpetration among School-Aged Children across 41 Countries: a Perspective of Routine Activity Theory," *Int. J. Bullying Prev.*, pp. 1–1, Jul. 2020, doi: 10.1007/s42380-020-00072-5.

[2] Q. Li, Y. Luo, Z. Hao, B. Smith, Y. Guo, and C. Tyrone, "Risk Factors of Cyberbullying Perpetration Among School-Aged Children Across 41 Countries: a Perspective of Routine Activity Theory," *Int. J. Bullying Prev.*, pp. 1–13, Jun. 2020, doi: 10.1007/s42380-020-00071-6.

[3] M. Hidajat, A. R. Adam, M. Danaparamita, and S. Suhendrik, "Dampak Media Sosial dalam Cyber Bullying," *ComTech Comput. Math. Eng. Appl.*, vol. 6, no. 1, p. 72, 2015, doi: 10.21512/comtech.v6i1.2289.

[4] S. L. Bryson, C. M. Brady, K. K. Childs, and K. Gryglewicz, "A Longitudinal Assessment of the Relationship Between Bullying Victimization, Symptoms of Depression, Emotional Problems, and Thoughts of Self-Harm Among Middle and High School Students," *Int. J. Bullying Prev.*, pp. 1–14, Jul. 2020, doi: 10.1007/s42380-020-00073-4.

[5] R. M. Kowalski *et al.*, "Hazing, Bullying, and Moral Disengagement," *Int. J. Bullying Prev.*, pp. 1–9, Jun. 2020, doi: 10.1007/s42380-020-00070-7.

[6] I. Markkanen, R. Välimaa, and L. Kannas, "Forms of Bullying and Associations Between School Perceptions and Being Bullied Among Finnish Secondary School Students Aged 13 and 15," *Int. J. Bullying Prev.*, pp. 1–10, Dec. 2019, doi: 10.1007/s42380-019-00058-y.

[7] Noviantho, S. M. Isa, and L. Ashianti, "Cyberbullying classification using text mining," in *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)*, Nov. 2017, pp. 241–246, doi: 10.1109/ICICOS.2017.8276369.

[8] N. Buslim, B. Busman, N. S. Sinatrya, and T. S. Kania, "Analisa Sentimen Menggunakan Data Twitter, Flume, Hive Pada Hadoop dan Java Untuk Deteksi Kemacetan di Jakarta," *J. Online Inform.*, vol. 3, no. 1, p. 1, 2018, doi: 10.15575/join.v3i1.141.

[9] N. Hayatin, G. I. Marthasari, and L. Nuarini, "Optimization of Sentiment Analysis for Indonesian Presidential Election using Naive Bayes and Particle Swarm Optimization," *J. Online Inform.*, vol. 5, no. 1, pp. 81–88, 2020.

[10] X. Bai, "Text classification based on LSTM and attention," in *2018 13th International Conference on Digital Information Management, ICDIM 2018*, Sep. 2018, pp. 29–32, doi: 10.1109/ICDIM.2018.8847061.

[11] B. B. Sahoo, R. Jha, A. Singh, and D. Kumar, "Long short-term memory (LSTM) recurrent neural network for low-flow hydrological time series forecasting," *Acta Geophys.*, vol. 67, no. 5, pp. 1471–1481, Oct. 2019, doi: 10.1007/s11600-019-00330-1.

[12] D. Soutner and L. Müller, "Application of LSTM neural networks in language modelling," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 8082 LNAI, pp. 105–112, doi: 10.1007/978-3-642-40585-3_14.

[13] N. K. Manaswi and N. K. Manaswi, "RNN and LSTM," in *Deep Learning with Applications Using Python*, Apress, 2018, pp. 115–126.

[14] J. Kim and N. Moon, "LSTM-Based Consumption Type Prediction Model," in *Lecture Notes in Electrical Engineering*, Dec. 2020, vol. 536 LNEE, pp. 564–567, doi: 10.1007/978-981-13-9341-9_97.

[15] W. Yu, X. Li, and J. Gonzalez, "Fast Training of Deep LSTM Networks," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Jul. 2019, vol. 11554 LNCS, pp. 3–10, doi: 10.1007/978-3-030-22796-8_1.

[16] S. D. Kumar and D. P. Subha, "Prediction of depression from EEG signal using long short term memory(LSTM)," in *Proceedings of the International Conference on Trends in Electronics and Informatics, ICOEI 2019*, Apr. 2019, vol. 2019-April, pp. 1248–1253, doi: 10.1109/icoei.2019.8862560.

[17] Y. Heryadi and H. L. H. S. Warnars, "Learning temporal representation of transaction amount for fraudulent transaction recognition using CNN, Stacked LSTM, and CNN-LSTM," in *2017 IEEE International Conference on Cybernetics and Computational Intelligence, CyberneticsCOM 2017 - Proceedings*, Mar. 2018, vol. 2017-November, pp. 84–89, doi: 10.1109/CYBERNETICSCOM.2017.8311689.

[18] M. Hajiaghayi and E. Vahedi, "Code failure prediction and pattern extraction using LSTM networks," in *Proceedings - 5th IEEE International Conference on Big Data Service and Applications, BigDataService 2019, Workshop on Big Data in Water Resources, Environment, and Hydraulic Engineering and Workshop on Medical, Healthcare, Using Big Data Technologies*, Apr. 2019, pp. 55–62, doi: 10.1109/BigDataService.2019.00014.

[19] S. Chakraborty, J. Banik, S. Addhya, and D. Chatterjee, "Study of Dependency on number of LSTM units for Character based Text Generation models," Jul. 2020, pp. 1–5, doi: 10.1109/iccsea49143.2020.9132839.

[20] H. Xue, D. Q. Huynh, and M. Reynolds, "SS-LSTM: A Hierarchical LSTM Model for Pedestrian Trajectory Prediction," in *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*, May 2018, vol. 2018-January, pp. 1186–1194, doi: 10.1109/WACV.2018.00135.

[21] A. Akandeh and F. M. Salem, "Slim LSTM NETWORKS: LSTM_6 and LSTM_C6," Jun. 2020, pp. 630–633, doi: 10.1109/mwscas.2019.8884912.

[22] Md Al-Amin, M. S. Islam, and S. Das Uzzal, "Sentiment analysis of Bengali comments with Word2Vec and sentiment information of words," in *ECCE 2017 - International Conference on Electrical, Computer and Communication Engineering*, Apr. 2017, pp. 186–190, doi: 10.1109/ECACE.2017.7912903.

[23] H. Tian and L. Wu, "Microblog Emotional Analysis Based on TF-IWF Weighted Word2vec Model," in *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, Mar. 2019, vol. 2018-November, pp. 893–896, doi: 10.1109/ICSESS.2018.8663837.

[24] J. Gao, Y. He, X. Zhang, and Y. Xia, "Duplicate short text detection based on Word2vec," in *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, Apr. 2018, vol. 2017-November, pp. 33–37, doi: 10.1109/ICSESS.2017.8342858.

[25] I. U. Ogul, C. Ozcan, and O. Hakdagli, "Keyword extraction based on word synonyms using WORD2VEC," in *27th Signal Processing and Communications Applications Conference, SIU 2019*, Apr. 2019, doi: 10.1109/SIU.2019.8806496.

[26] M. S. R. Hitesh, V. Vaibhav, Y. J. A. Kalki, S. H. Kamtam, and S. Kumari, "Real-time sentiment analysis of 2019 election tweets using word2vec and random forest model," in *2019 2nd International Conference on Intelligent Communication and Computational Techniques, ICCT 2019*, Sep. 2019, pp. 146–151, doi: 10.1109/ICCT46177.2019.8969049.

[27] W. Tian, J. Li, and H. Li, "A Method of Feature Selection Based on Word2Vec in Text Categorization," in *Chinese Control Conference, CCC*, Oct. 2018, vol. 2018-July, pp. 9452–9455, doi: 10.23919/ChiCC.2018.8483345.

[28] M. Kim, J. Kim, K. Kim, and M. Sohn, "Ontology-aided Word2vec based synonym identification for ontology alignment," in *Proceedings - 2020 IEEE International Conference on Big Data and Smart Computing, BigComp 2020*, Feb. 2020, pp. 429–431, doi: 10.1109/BigComp48618.2020.00-35.

[29] M. Tezgider, B. Yildiz, and G. Aydin, "Improving Word Representation by Tuning Word2Vec Parameters with Deep Learning Model," in *2018 International Conference on Artificial Intelligence and Data Processing, IDAP 2018*, Jan. 2019, doi: 10.1109/IDAP.2018.8620919.

978-1-7281-7596-6/20/$31.00 ©2020 IEEE

Deep Learning Approach for Bullying Classification on Twitter Social Media with Indonesian Language

ORIGINALITY REPORT

18%	10%	10%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

4%

★ Submitted to Coventry University

Student Paper

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		