JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 10. NO. 1 AUGUST 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i1.5239

MODEL OF INDONESIAN CYBERBULLYING TEXT DETECTION USING MODIFIED LONG SHORT-TERM MEMORY

Mariana Purba[1*]; Paisal[2]; Cahyo Pambudi Darmo[3]; Handrie Noprisson[4]; Vina Ayumi[5]

Department of Informatics[1,2,3]
Universitas Sjakhyakirti, Indonesia[1,2,3]
https://unisti.ac.id[1,2,3]
mariana_purba@unisti.ac.id[1*], paisal@unisti.ac.id[2], cahyopambudi@unisti.ac.id[3]

Department of Informatics Engineering[4,5]
Universitas Dian Nusantara, Indonesia[4,5]
https://undira.ac.id/[4,5]
handrie.noprisson@dosen.undira.ac.id[4], vina.ayumi@dosen.undira.cid[5]

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract—*Cyberbullying, in its essence, refers to the deliberate act of exploiting technological tools to inflict harm upon others. Typically, this offensive conduct is perpetuated repeatedly, as the perpetrator takes solace in concealing their true identity, thereby avoiding direct exposure to the victim's reactions. It is worth noting that the actions of the cyberbully and the responses of the individual being cyberbullied share an undeniable interconnection. The main objective of this study was to identify and analyze Instagram comments that contain bullying words using a model of WLSTML2 which is an optimization of a long short-term memory network with word-embedding and L2 regularization. This experiment using dataset with negative labels as many as 400 data and positive as many as 400 data. In this study, a comparison of 70% training data and 30% testing data was used. Based on experimental results, the WLSTMDR model obtained 100% accuracy at the training stage and 80% accuracy at the testing stage. The WLSTML2 model received an accuracy of 99.25% at the training stage and an accuracy of 83% at the testing stage. The WLSTML1 model obtained an accuracy of 97.01% at the training stage and an accuracy of 80% at the testing stage. Based on the experimental results, the WLSTML2 model gets the best accuracy at the training and testing stages. At the testing stage of 132 data, it was found that the positive label data predicted to be correct was 56 data and the negative label data that was predicted to be correct was 53 data.*

Keywords: *cyberbullying, embedding layer, LSTM, regularization*

Intisari—*Cyberbullying, mengacu pada tindakan yang disengaja mengeksploitasi teknologi untuk menimbulkan kerugian pada orang lain. Biasanya, perilaku ini dilakukan berulang kali, karena pelaku dapat menyembunyikan identitas asli, sehingga menghindari reaksi langsung dari korban. Tindakan cyberbully dan reaksi individu yang mengalami cyberbullying dapat berkelanjutan dan menyebabkan permasalahan di dunia nyata. Tujuan utama dari penelitian ini adalah untuk mengidentifikasi dan menganalisis komentar Instagram yang mengandung kata-kata bullying menggunakan model WLSTML2 yaitu optimalisasi jaringan memori jangka pendek panjang dengan word-embedding dan regularisasi L2. Penelitian ini menggunakan dataset dengan label negatif sebanyak 400 data dan positif sebanyak 400 data. Dalam eksperimen ini, perbandingan data pelatihan 70% dan data pengujian 30% digunakan. Berdasarkan hasil penelitian, model WLSTMDR diperoleh akurasi 100% pada tahap pelatihan dan akurasi 80% pada tahap pengujian. Model WLSTML2 mendapatkan akurasi 99,25% pada tahap pelatihan dan akurasi 83% pada tahap pengujian sedangkan model WLSTML1 memperoleh akurasi 97,01% pada tahap pelatihan dan akurasi 80% pada tahap pengujian. Model WLSTML2 mendapatkan akurasi terbaik pada tahap pelatihan dan pengujian. Pada tahap pengujian*

terhadap 132 data menggunakan WLSTML2, diketahui bahwa data label positif yang diprediksi benar adalah 56 data dan data label negatif yang diprediksi benar adalah 53 data.

***Kata Kunci**: perundungan siber, lapisan embedding, LSTM, regulasi*

INTRODUCTION

Cyberbullying is the act of bullying through technological devices to intentionally hurt others. This action is usually done repeatedly because the perpetrator feels safe by hiding his identity so that he does not get to see the victim's response directly. The behavior of the cyberbully and the behavior of the victim of cyberbullying are related. The more reactive the behavior of the perpetrator, the more reactive the behavior of the victim [1]–[3].

The main objective of this study was to identify and analyze Instagram comments that contain bullying meaning, with a particular focus on interactions on social media. This analysis is used to distinguish Instagram comments that are negative or positive so as to provide a clearer picture of how Instagram is developing new features in filtering Instagram comments that contain the meaning of bullying. More detailed research into Instagram comments is expected to make a significant contribution to the development of more effective bullying prevention strategies on the Instagram platform [4]–[7].

A long short-term memory (LSTM) model has been proposed for the detection and prevention of cyberbullying. The model uses a deep learning approach to extract features, train the model, and analyze the data, producing more accurate results compared to previous machine learning models. The proposed LSTM model achieves an accuracy of 75.12%, which is significantly better than the accuracy achieved by other models [8]–[12].

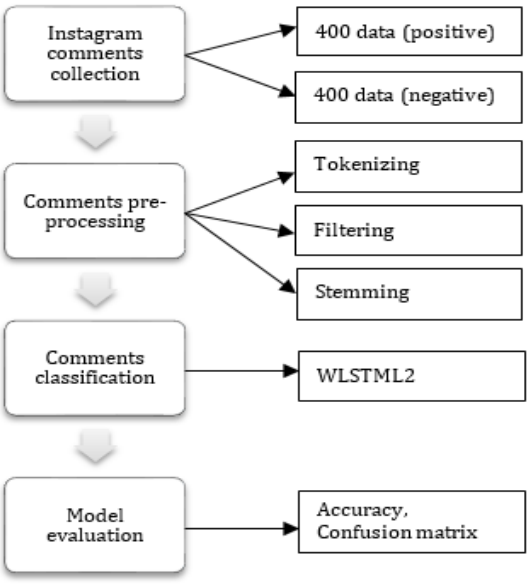
This optimization of the LSTM-based model contributes to the effective identification of cyberbullying attacks and on social media platforms [13], [14]. Optimization of LSTM can be done by adding an embedding layer. The use of embedding layers on LSTM aims to extract and classify key semantic information in text. Several studies discuss the use of embedding layers on LSTM in text classification and compare the effectiveness of word embedding techniques in text classification using LSTM [15]–[19].

In addition, another problem of LSTM-based models for text classification is the problem of overfitting, which can lead to low prediction accuracy. To address this issue, researchers have proposed various techniques such as L1 and L2 regularization and dropout. Dropout regularization randomly drops units (neurons) from the neural network during training, preventing units from depending too much on each other and reducing overfitting. L1 and L2 regularization can also be used to control the complexity of the model and prevent overfitting. By using these techniques, LSTM have achieved improved prediction accuracy and reduced overfitting in various applications, including text classification [20]–[25].

This study used the LSTM algorithm for the classification of Instagram comments. The LSTM algorithm has advantages in learning patterns in Instagram comments and can model the weighted content of the network so that the process of classifying Instagram comments will be easier. The aim of this research is proposed the model of WLSTML2, WLSTML1, WLSTMDR and which is optimization of long short-term memory network with word-embedding and L2, L1 or dropout regularization for cyberbullying detection in Indonesian text.

MATERIALS AND METHODS

The research methodology is divided into four phases. The phases are Instagram comments collection, comments pre-processing, comments classification and model evaluation. Every phase has different techniques of method to accomplish the goal. The detail of the research methodology is depicted in Figure 1.



Source: (Research Results, 2024)

Figure 1. Research methodology

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 1 AUGUST 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i1.5239

The dataset used in this study is secondary data. The amount of data used was 400 comments from Instagram users. Data in the form of comments from Instagram by scraping using application programming interface services that have been processed in previous studies. The process of the positive or negative label based on keywords used in retrieving comments are based on words that contain the meaning of mocking or vilifying an object. Labeling each comment is done by giving two classes, namely positive class, negative class. Negative class means Instagram comments that contain the meaning of bully and positive class means Instagram comments contain meaning not bully (tends to the meaning of motivation or support).

In the second stage, there are three processes, namely tokenization, filtering and stemming. The tokenization process is carried out by cutting Instagram comments based on space characters into several pieces based on each word that makes up a comment. The result of tokenization called a token is a single word that will characterize the classification of Instagram comments. The second process is filtering. This process is carried out by taking important words from the results of the tokenization process. In this process, unimportant words (stop-words) will be eliminated to reduce the number of words that will be processed next. The third process is stemming. This process is done by converting affixes from filtered words to stem (root words).

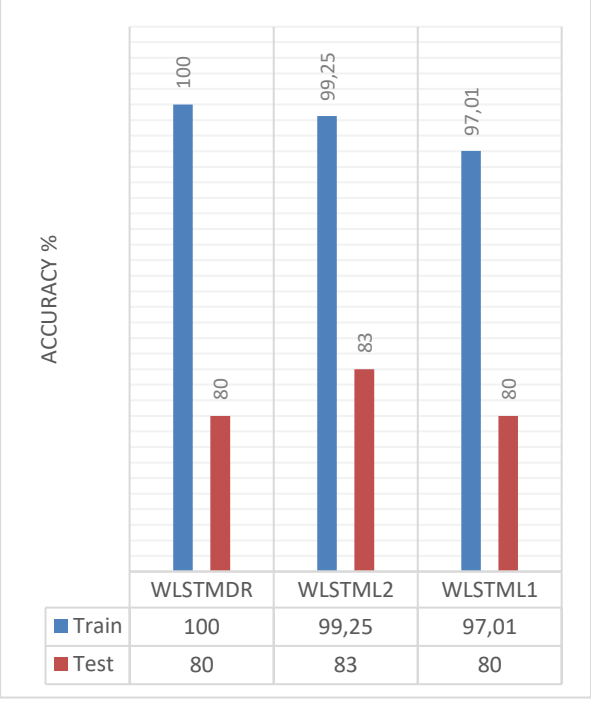
The third stage is the use of the LSTM method to conduct an analysis of Instagram comment classification. LSTM is one of the developments of RNN which has an inability to store information during the learning process if too much information must be stored. LSTM is used to replace RNN nodes in the hidden layer which are also referred to as LSTM cells. In LSTM will be implemented a dropout layer and Adam optimizer. Evaluation of Instagram comment classification is done by comparing between prediction data and actual data. Prediction data is in the form of comment classification results generated by the LSTM algorithm while actual data is in the form of comment classification results generated from manual labeling. In this study, the evaluation used is accuracy by comparing cases that are classified correctly with the number of all existing classification cases.

RESULTS AND DISCUSSION

This section describes the results of experiments to analyze data for negative labels as many as 400 data and positive as many as 400 data. Next, the data is split to avoid overfitting. The splitting method is done based on labels, then the data is recombined into data train and test. In this study, a comparison of 70% training data and 30% testing data was used.

The next stage is experiments for data classification. In this study, no LSTM architectural model was compared which was optimized with layer embedding and regularization of the output layer. Layer embedding to convert text input into vector representation while regularization on layer output is used to reduce overfitting. The first model is WLSTMDR. This model is an optimization of a long short-term memory network with word-embedding and dropout regularization. The second model is WLSTML2 which is an optimization of a long short-term memory network with word-embedding and L2 regularization. The last model is WLSTML1 which is an optimization of long short-term memory network with word-embedding and L1 regularization.

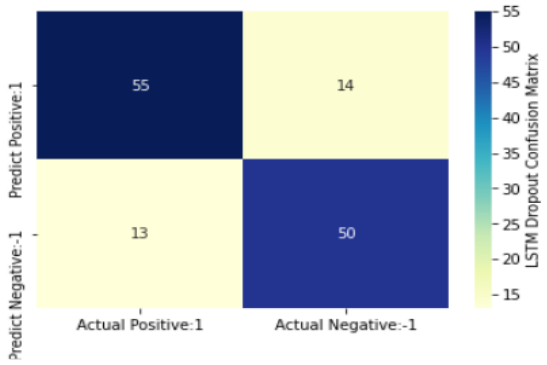
Based on experimental results, the WLSTMDR model obtained 100% accuracy at the training stage and 80% accuracy at the testing stage. The WLSTML2 model received an accuracy of 99.25% at the training stage and an accuracy of 83% at the testing stage. The WLSTML1 model obtained an accuracy of 97.01% at the training stage and an accuracy of 80% at the testing stage. The comparison results of each model can be seen in Figure 2.



Source: (Research Results, 2024)

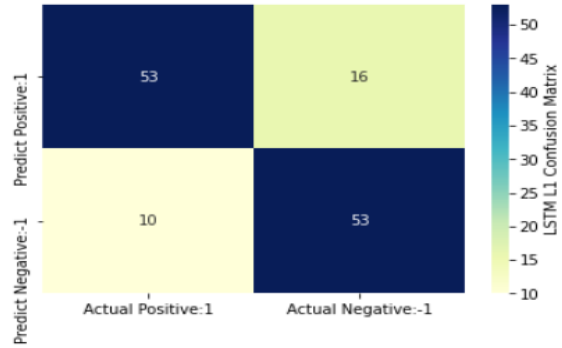
Figure 2. Comparison of experiment result

Analysis of experimental results at the testing stage using the WLSTMDR model is also presented in the form of a confusion matrix. In the testing phase of 132 data, it was known that the positive label data predicted to be correct was 55 data and the negative label data that was predicted to be correct was 50 data. The distribution of prediction data using WLSTMDR at the testing stage can be seen in the confusion matrix in **Figure 3**.



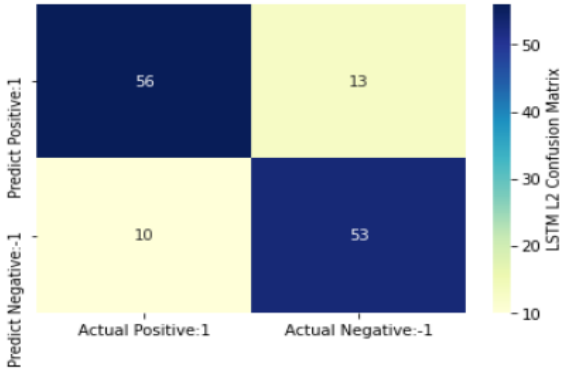
Source: (Research Results, 2024)
Figure 3. Confusion matrix of WLSTMDR

Analysis of experimental results at the testing stage using WLSTML1 model is also presented in the form of a confusion matrix. In the testing phase of 132 data, it was found that the positive label data predicted to be correct was 53 data and the negative label data that was predicted to be correct was 53 data. If calculated using the accuracy formula, WLSTML1 model obtained an accuracy of 80% at the testing stage. The distribution of prediction data using WLSTML1 at the testing stage can be seen in the confusion matrix in **Figure 4**.



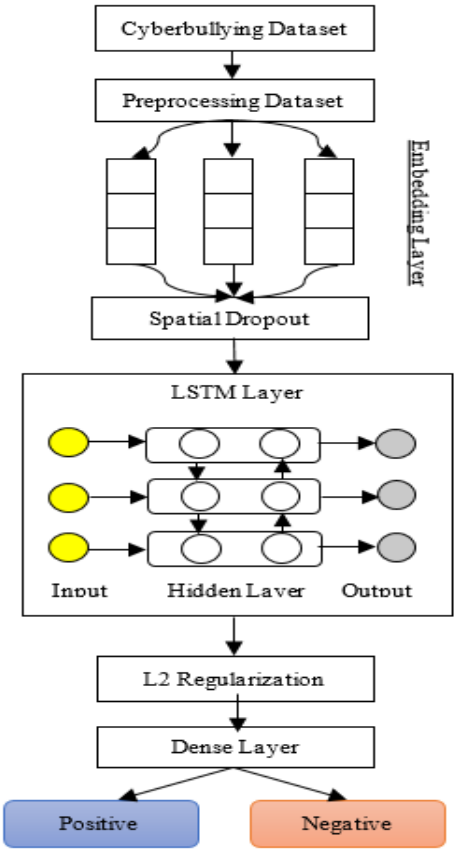
Source: (Research Results, 2024)
Figure 4. Confusion matrix of WLSTML1

Based on the experimental results, the WLSTML2 model gets the best accuracy at the training and testing stages. At the testing stage of 132 data, it was found that the positive label data predicted to be correct was 56 data and the negative label data that was predicted to be correct was 53 data. If calculated using the accuracy formula, The WLSTML2 model obtained an accuracy of 83% at the testing stage. The distribution of prediction data using WLSTML2 at the testing stage can be seen in the confusion matrix in **Figure 5**.



Source: (Research Results, 2024)
Figure 5. Confusion matrix of WLSTML2

Based on the results of the analysis of the accuracy and confusion matrix, the WLSTML2 model gets the best accuracy at the training and testing stages. The WLSTML2 model received an accuracy of 99.25% at the training stage and an accuracy of 83% at the testing stage. The WLSTML2 architecture used in this study can be seen in Figure 6.



Source: (Research Results, 2024)
Figure 6. Architecture of WLSTML2

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 1 AUGUST 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i1.5239

The architecture WLSTML2 is an LSTM-based model using multiple layers to process text input. First, the embedding layer converts the input text into a vector. The utilization of embedding layers on long short-term memory (LSTM) endeavors to extract and categorize essential semantic information within textual data. Numerous investigations delve into the employment of embedding layers on LSTM for text classification purposes, while also assessing the efficacy of word embedding in text classification through the utilization of LSTM.

In addition, WLSTML2 using regularization method to tackle additional overfitting issue with text classification models based on LSTM, which can result in diminished accuracy of predictions. In order to tackle this concern, several techniques including dropout, L1 and L2 regularization, are compared to obtained best accuracy. Dropout regularization entails randomly excluding units (neurons) from the neural network during the training process, thereby preventing excessive reliance on interconnected units and mitigating overfitting. Additionally, L1 and L2 regularization can be employed to govern the complexity of the model and avert overfitting. By incorporating these techniques, long short-term memory (LSTM) models have successfully demonstrated enhanced prediction accuracy and reduced overfitting across various applications, encompassing text classification. However, this research doesn't handle the challenges of sarcasm, slang, or misspellings in the comments and will applied in next experiment.

CONCLUSION

The utilized dataset consisted of 400 comments derived from Instagram users. The data was collected in the form of comments obtained from Instagram through the utilization of application programming interface services. In this particular investigation, a comparative analysis was conducted using a split of 70% for training data and 30% for testing data. Upon examination of the experimental outcomes, it was revealed that the WLSTMDR model achieved a commendable accuracy rate of 100% during the training phase, while exhibiting an accuracy rate of 80% during the testing phase. Similarly, the WLSTML2 model demonstrated a notable accuracy rate of 99.25% during the training phase, and a slightly lower accuracy rate of 83% during the testing phase. The WLSTML1 model, on the other hand, attained an accuracy rate of 97.01% during the training stage, with a corresponding accuracy rate of 80% during the testing stage. In light of the experimental findings, it is evident that the WLSTML2 model performed most admirably, surpassing the other models in terms of accuracy during both the training and testing stages. Specifically, when examining the testing stage encompassing a dataset of 132 data, it was identified that 56 data of positively labeled data were accurately predicted, whereas 53 data of negatively labeled data were correctly predicted.

REFERENCE

[1] G. R., M. G., and D. M. S. Anbarasi, "Detection and Classification of Cyberbullying Using CR*," *Int. J. Sci. Technol. Eng.*, vol. 11, no. 4, pp. 24–29, pp. 24-29, Apr. 2023, doi: 10.22214/ijraset.2023.49984.

[2] K. Rong, X.-W. Chu, and Y. Zhao, "Qualitative analyses on the classification model of bystander behavior in cyberbullying," *Frontiers in psychology,* vol. 14, p. 1152331, Jul. 2023, doi: 10.3389/fpsyg.2023.1152331.

[3] S. Perumal, "A Survey on Cyberbullying Classification and Detection," *J. Inf. Technol. Digit. World*, vol. 5, no. 2, pp. 85–92, 2023, doi: 10.36548/jitdw.2023.2.001.

[4] M. Alauthman, S. R. Yonbawi, and A. Almomani, "Cyberbullying Detection and Recognition with Type Determination Based on Machine Learning," *Computers, Materials & Continua*, vol. 75, no. 3, pp. 5307–5319, 2023, doi: 10.32604/cmc.2023.031848.

[5] N. Jahan and R. H. Nabil, "Machine Learning in Cyberbullying Detection from Social-Media Image or Screenshot with Optical Character Recognition," *I.J. Intelligent Systems and Applications*, vol. 15, no. 2, pp. 1–13, Apr. 2023, doi: 10.5815/ijisa.2023.02.01

[6] D. H. and M. Manimaran., "A Review of Machine Learning and AI-Based Approaches to Detecting Cyberbullying on Social Media," *Int. J. Sci. Technol. Eng.*, vol. 11, no. 4, pp. 1594–1602, Apr. 2023, doi: 10.22214/ijraset.2023.50438.

[7] S. D. and R. K. Dandu, "Machine Learning and Deep Learning Algorithm for Online Bullying Identification," *Int. J. Sci. Technol. Eng.*, vol. 11, no. 6, pp. 2708–2711, Jun. 2023, doi: 10.22214/ijraset.2023.53951.

[8] Abinaya, K., Jayakumar, D., & Sneha, S, "Bi-LSTM Neural Network Approach to Detect and Recognize Cyberthreats, Cyberstalking and Extremist Tweets in Twitter," in *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, May. 2023, pp. 1286–1290, doi: 10.1109/ICAAIC56838.2023.10140281.

[9] S. A. Kahate and A. D. Raut, "Design of a Deep Learning Model for Cyberbullying and Cyberstalking Attack Mitigation via Online Social Media Analysis," in *2023 4th International Conference on Innovative Trends in Information Technology (ICITIIT)*, pp. 1–7, Feb. 2023, doi: 0.1109/ICITIIT57246.2023.10068711

[10] M. A. Arsha and D. K. Daniel, "Cyberbullying Detection on Social Networks using LSTM Model," in *2022 International Conference on Innovations in Science and Technology for Sustainable Development (ICISTSD)*, pp. 293–296, Aug. 2022, doi: 10.1109/ICISTSD55159.2022.10010559

[11] M. Purba, E. Ermatita, A. Abdiansah, V. Ayumi, H. Noprisson, and A. Ratnasari, "A Systematic Literature Review of Knowledge Sharing Practices in Academic Institutions," in *2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS*, pp. 337–342, Oct. 2021, doi: 10.1109/ICIMCIS53775.2021.9699350.

[12] M. Purba *et al.*, "Effect of Random Splitting and Cross Validation for Indonesian Opinion Mining using Machine Learning Approach," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 9, 2022, doi: 10.14569/IJACSA.2022.0130917

[13] M. T. Hasan, M. A. E. Hossain, M. S. H. Mukta, A. Akter, M. Ahmed, and S. Islam, "A review on deep-learning-based cyberbullying detection," *Futur. Internet*, vol. 15, no. 5, p. 179, 2023, doi: 10.3390/fi15050179.

[14] M. Arif, "A systematic review of machine learning algorithms in cyberbullying detection: future directions and challenges," *J. Inf. Secur. Cybercrimes Res.*, vol. 4, no. 1, pp. 1–26, Jun. 2021, doi: 10.26735/GBTV9013.

[15] M. R. Ilham and A. D. Laksito, "Comparative Analysis of Using Word Embedding in Deep Learning for Text Classification," *J. Ris. Inform.*, vol. 5, no. 2, pp. 195–202, Mar. 2023, doi: 10.34288/jri.v5i2.208.

[16] P. Li, Y. Liu, Y. Hu, Y. Zhang, X. Hu, and K. Yu, "A Drift-Sensitive Distributed LSTM Method for Short Text Stream Classification," *IEEE Trans. Big Data*, vol. 9, pp. 341–357, Apr. 2023, doi: 10.1109/TBDATA.2022.3164239.

[17] V. Ayumi and I. Nurhaida, "Prediction Using Markov for Determining Location of Human Mobility," *Int. J. Inf. Sci. Technol. – iJIST*, vol. 4, no. 1, pp. 1–6, Feb. 2020, doi: dx.doi.org/10.57675/IMIST.PRSM/ijist-v4i1.141.

[18] M. Sadikin and A. Fauzan, "Evaluation of Machine Learning Approach for Sentiment Analysis using Yelp Dataset," *Eur. J. Electr. Eng. Comput. Sci.*, vol. 7, no. 6, pp. 58–64, Dec. 2023, doi: 10.24018/ejece.2023.7.6.583.

[19] H. Noprisson, "Evaluation of Information System Implementation Support for 6-Area Smart City Development," *JSAI (Journal Sci. Appl. Informatics)*, vol. 6, no. 1, pp. 83–88, Jan. 2023, doi: 10.36085/jsai.v6i1.6087

[20] W. Yang, C. Jia, and R. Liu, "Construction and Simulation of the Enterprise Financial Risk Diagnosis Model by Using Dropout and BN to Improve LSTM," *Secur. Commun. Networks*, vol. 2022, pp. 1–9, 2022, doi: 10.1155/2022/4767980.

[21] X. Xie, M. Xie, A. J. Moshayedi, and M. H. N. Skandari, "A Hybrid Improved Neural Networks Algorithm Based on L2 and Dropout Regularization," *Math. Probl. Eng.*, vol. 2022, pp. 1–19, 2022, doi: 10.1155/2022/8220453.

[22] B. Pansambal, "Integrating Dropout Regularization Technique at Different Layers to Improve the Performance of Neural Networks," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 4, pp. 716-722, 2023, doi: 10.14569/IJACSA.2023.0140478.

[23] H. Noprisson, E. Ermatita, A. Abdiansah, V. Ayumi, M. Purba, and H. Setiawan, "Fine-Tuning Transfer Learning Model in Woven Fabric Pattern Classification," *Int. J. Innov. Comput. Inf. Control*, vol. 18, no. 06, p. 1885, 2022, doi: 10.24507/ijicic.18.06.1885.

[24] D. Ramayanti *et al.*, "Tuberculosis Ontology Generation and Enrichment Based Text Mining," in *2020 International Conference on Information Technology Systems and Innovation (ICITSI)*, pp. 429–434, 2020, doi: 10.1109/ICITSI50517.2020.9264922.

[25] Y. Jumaryadi, D. Firdaus, B. Priambodo, and Z. P. Putra, "Determining the Best Graduation Using Fuzzy AHP," in *2020 2nd International Conference on Broadband Communications, Wireless Sensors and Powering (BCWSP)*, pp. 59–63, 2020, doi: 10.1109/BCWSP50066.2020.9249463.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri