This could happen because the IndoBERTweet model was trained using Twitter data by crawling Indonesian tweets using the official Twitter API.

Therefore, the BiLSTM layer helps the model to perform better. This was not the case with the CNN layer, as in our research, adding the CNN layer only improved the performance a bit, and the fine-tuned model still performed better than adding the CNN layer. Using the first dataset, our combined model with the BilSTM layer improved the performance slightly more than the fine-tuned IndoBERTweet. However, the second dataset significantly improved by 6% from the IndoBERTweet model and 5% increased performance from fine-tuned IndoBERTweet model.

When training our model, we used a small number of the epoch because the size of our dataset is relatively small and imbalanced. This small number of epochs helps the model not overfit; additionally, we add a dropout layer to reduce the overfitting. We also tried using epoch with large numbers, such as 10, 15, and 20. As a result, our model training accuracy keeps improving while validation accuracy could not keep up with training accuracy. Most of the time, the validation accuracy went up and down until our model finished training.

IV. Conclusion

The combined model has been able to classify hate speech properly through text processing, such as cleaning and converting slang and misspelled words to their original form and tokenizing the sentence. The token gained from tokenizer vocabulary using BertTokenizer and fed into the model. The result obtained from the model is evaluated using a confusion matrix. Based on the analysis, the model composed of IndoBERTweet and BiLSTM obtained much better results with the highest score of accuracy, recall, precision, and F1 score are 93.7%, 92.9%, 93.8%, and 93.3%, respectively. The model with IndoBERTweet and CNN, meanwhile, gained the lowest result with accuracy, recall, precision, and F1 scores of 88%, 87.7%, 87.5%, and 87.6%, respectively. Although the model used the CNN layer, there is no significant increase in performance gained from only using IndoBERTweet. There is also a slight difference between the results using different datasets. The difference varies between 2% to 5%, and the data quality, such as the type of word and the cleaning process, played a decisive factor in this matter. It is hoped that this research could lead to more research that focuses on datasets. Another use of other RNN layer is also a consideration to increase the performance of hate speech classification further.

References

[1] S. Kemp, "Digital 2021: Global Overview Report," 27 January 2021. [Online]. Available: https://datareportal.com/reports/digital-2021-global-overview-report. [Accessed 25 Maret 2021].

[2] W. Warner and J. Hirschberg, "Detecting Hate Speech on the World Wide Web," *Proceedings of the second workshop on language in social media,* pp. 19-26, 2012.

[3] J. W. Howard, "Free speech and hate speech," *Annual Review of Political Science,* pp. 93-109, 2019.

[4] B. Mathew, R. Dutt, P. Goyal and A. Mukherjee, "Spread of Hate Speech in Online Social Media," *Proceedings of the 10th ACM Conference on Web Science - WebSci '19,* pp. 173-182, 2019.

[5] H. I. Harahap, "Hate Speech in Election: Increasing Trends and Concerns," *Advances in Social Science, Education and Humanities Research,* pp. 44-46, 2019.

[6] A. Purnomo, "Legal perspectives concerning hate speech in indonesia," *PalArch's Journal of Archaeology of Egypt/Egyptology,* pp. 544-554, 2020.

[7] M. Mozafari, R. Farahbakhsh and N. Crespi, "A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media," *COMPLEX NETWORKS 2019, SCI 881,* pp. 928-940, 2020.

[8] G. B. Herwanto, A. M. Ningtyas, I. G. Mujiyatna, I. N. P. Trisna and K. E. Nugraha, "Hate Speech Detection in Indonesian Twitter using Contextual Embedding Approach," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems),* pp. 177-188, 2021.

[9] F. Koto, J. H. Lau and T. Baldwin, "INDOBERTWEET: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," *In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021),* pp. 1-9, 2021.

[10] I. Alfina, R. Mulia, M. I. Fanany and Y. Ekanata, "Hate Speech Detection in the Indonesian Language: A Dataset and Preliminary Study," *Proceeding of 9th International Conference on Advanced Computer Science and Information Systems 2017(ICACSIS 2017),* pp. 233-238, 2017.

[11] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," *Proceedings of the Third Workshop on Abusive Language Online,* pp. 46-57, 2019.

[12] F. D. Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi and M. Tesconi, "Hate me, hate me not: Hate speech detection on Facebook," *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17),* pp. 86-95, 2017.

[13] B. Gambäck and U. K. Sikdar, "Using Convolutional Neural Networks to Classify Hate-Speech," *Proceedings of the First Workshop on Abusive Language Online,* pp. 85-90, 2017.

[14] Z. Al-Makhadmeh and A. Tolba, "Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach," *Computing,* pp. 501-522, 2019.

[15] H. Faris, I. Aljarah, M. Habib and P. A. Castillo, "Hate Speech Detection using Word Embedding and Deep Learning in the Arabic Language Context," *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2020),* pp. 453-460, 2020.

[16] P. Kapil, A. Ekbal and D. Das, "Investigating Deep Learning Approaches for Hate Speech Detection in Social Media," unpublished, 2020.

[17] T. B. Nguyen, Q. M. Nguyen, T. H. Nguyen, N. P. Pham, T. L. Nguyen and Q. T. Do, "VAIS Hate Speech Detection System: A Deep Learning based Approach for System Combination," unpublished, 2019.

[18] S. S. Aluru, B. Mathew, P. Saha and A. Mukherjee, "Deep Learning Models for Multilingual Hate Speech Detection," unpublished, 2020.

[19] T. L. Sutejo and D. P. Lestari, "Indonesia Hate Speech Detection using Deep Learning," *International Conference on Asian Language Processing (IALP),* pp. 39-43, 2018.

[20] A. R. Isnain, A. Sihabuddin and Y. Suyanto, "Bidirectional Long Short Term Memory Method and Word2vec Extraction Approach for Hate Speech Detection," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems),* pp. 169-178, 2020.

[21] A. Marpaung, R. Rismala and H. Nurrahmi, "Hate Speech Detection in Indonesia Twitter Texts using Bidirectional Gated Recurrent Unit," *International Conference on Knowledge and Smart Technology (KST),* no. 13, pp. 186-190, 2021.

[22] R. Aggarwal, "Bi-LSTM," 4 July 2019. [Online]. Available: https://medium.com/@raghavaggarwal0089/bi-lstm-bc3d68da8bd0. [Accessed 12 August 2021].

[23] V. Kotu and B. Deshpande, Data Science (Second Edition), Elsevier, 2019.