IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Impact Factor: 6.252

Volume 2, Issue 1, November 2022

Expeditious Cyber-Bullying Detection Method Utilizing Compact BERT Models

Rushikesh Ambre[1], Yogendra Yadav[2], Nikhil Kamble[3], Gaytri Khule[4], Dr. Sunil Khatal[5]

Students, Department of Computer Engineering[1,2,3,4]
Professor, Department of Computer Engineering[5]
Sharadchandra Pawar College of Engineering, Pune, Maharashtra, India

Abstract: *Since many people now use social media to disseminate hate, many researchers have been concentrating on the issue of detecting cyberbullying. in the last ten years. In this work, transfer learning is used to address this issue. We adjust our tiny BERT models using data from hate speech. We use the Focal Loss function to address the data's class imbalance. With this method, we were able to obtain cutting-edge outcomes on the hate speech dataset, including 0.91 precision, 0.92 recall, and 0.91 F1-score. Additionally, we demonstrate using our transfer learning pipeline that the more compact BERT models are much faster at detecting cyberbullying and are appropriate for real-time applications.*

Keywords: Focal Loss, Transfer Learning, Hate Speech, Compact BERT, Cyber bullying

REFERENCES

[1].M. P. Hamm, A. S. Newton, A. Chisholm, J. Shulhan, A. Milne, P. Sundar, H. Ennis, S. D. Scott, and L. Hartling, “Prevalence and effect of cyberbullying on children and young people: A scoping review of social media studies,” JAMA pediatrics, vol. 169, no. 8, pp. 770–777, 2015.

[2].S. Salawu, Y. He, and J. Lumsden, “Approaches to automated detection of cyberbullying: A survey,” IEEE Transactions on Affective Computing, 2017.

[3].I. Turc, M.-W. Chang, K. Lee, and K. Toutanova, “Well-read students learn better: On the importance of pre-training compact models,” arXiv preprint arXiv:1908.08962v2, 2019.

[4].M. Dadvar and F. De Jong, “Cyberbullying detection: a step toward a safer internet yard,” in Proceedings of the 21st International Conference on World Wide Web, 2012, pp. 121–126

[5].A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards, “Detecting cyberbullying: query terms and techniques,” in Proceedings of the 5th annual acm web science conference, 2013, pp. 195–204. [6] P. Singh and S. Chand, “Pardeep at semeval-2019 task 6: Identifying and categorizing offensive language in social media using deep learning,” in Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 727–734.

[6]. V. Golem, M. Karan, and J. Snajder, “Combining shallow and deep ˇ learning for aggressive text detection,” in Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), 2018, pp. 188–198.

[7]. A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, “A unified deep learning architecture for abuse detection,” in Proceedings of the 10th ACM Conference on Web Science, 2019, pp. 105–114.

[8]. T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.

[9]. J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[10]. M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, “Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval),” arXiv preprint arXiv:1903.08983, 2019.

[11]. P. Liu, W. Li, and L. Zou, “Nuli at semeval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers,” in Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 87–91.

DOI: 10.48175/IJARSCT-7379

[12]. P. Aggarwal, T. Horsmann, M. Wojatzki, and T. Zesch, "Ltl-ude at semeval-2019 task 6: Bert and two-vote classification for categorizing offensiveness," in Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 678–682.

[13]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[14]. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in neural information processing systems, 2017, pp. 5998–6008.

[15]. A.-M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of twitter abusive behavior," in 11th International Conference on Web and Social Media, ICWSM 2018. AAAI Press, 2018.

[16].S. Srivastava, P. Khurana, and V. Tewari, "Identifying aggression and toxicity in comments using capsule network," in Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), 2018, pp. 98–105.

[17]. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss ´ for dense object detection," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[18]. L. Luo, Y. Xiong, Y. Liu, and X. Sun, "Adaptive gradient methods with dynamic bound of learning rate," arXiv preprint arXiv:1902.09843, 2019.

DOI: 10.48175/IJARSCT-7379