Cyberbullying Detection

Violeta-Nicoleta BĂDIȚĂ
Faculty of Electronics, Telecommunications and Information Technology,
University POLITEHNICA of Bucharest
violeta.badita@stud.etti.upb.ro

Abstract

The harmful effects of cyberbullying, including mental health problems, poor academic performance, and suicidal thoughts, highlight the importance of developing effective detection systems. This text discusses various approaches to detecting and combating cyberbullying on social media platforms. A cyberbullying detection system (CDS) could identify different types of bullying based on gender, religion, ethnicity, age, aggression, and non-cyberbullying. The CDS system utilized a hybrid deep learning architecture that integrated convolutional neural networks (CNN) with bidirectional long short-term memory networks (BiLSTM). Both binary and multiclass classification datasets were used, and the BiLSTM outperformed the combined CNN-BiLSTM classifier in detecting online bullying with an accuracy rate of 99%. A novel algorithm called CNN-CB was proposed to eliminate the need for feature engineering and to produce better predictions than traditional cyberbullying detection approaches. CNN-CB utilizes convolutional neural networks and incorporates semantics through the use of word embedding. Experiments showed that CNN-CB outperformed traditional content-based cyberbullying detection with an accuracy of 95%.

Index terms: cyberbullying, CNN-CB, BiLSTM, hybrid, BERT, DistilBERT

References

[1]. V. Nahar, X. Li and C. Pang, "An Effective Approach for Cyberbullying Detection", May 2013.

[2]. L. Zhang, S. Wang, and B. Liu, “Deep learning for sentiment analysis: A survey”, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018.

[3]. Monirah Abdullah Al-Ajlan, “Deep Learning Algorithm for Cyberbullying Detection”, IJACSA 2018f.

[4]. Dr. Vijayakumar V., Dr. Hari Prasad D., Adolf P., “Multimodal Cyberbullying Detection using Hybrid Deep Learning Algorithms”, 2021.

[5]. Theyazn H. H. Aldhyani, Mosleh Hmoud Al-Adhaileh and Saleh Nagi Alsubari, “Cyberbullying Identification System Based Deep Learning Algorithms”, 2022.

[6]. Aditya Desai, Shashank Kalaskar, Omkar Kumbhar , and Rashmi Dhumal, “Cyber Bullying Detection on Social Media using Machine Learning”, 2021.

[7]. Saranyanath Kadamgode Puthenveedu, “Cyberbullying Detection using Ensemble Method, Ontario”, Canada April 2022.

[8]. Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, “A neural probabilistic language model,” J. Mach. Learn. Res., vol. 3.

[9]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”.

[10]. "tf-idf: A single Page Tutorial," [Online]. Available: http://www.tfidf.com.

[11]. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”.

[12]. M. R. Costa-jussà, E. Gonzalez, A. Moreno, and E. Cumalat, “Abusive language in Spanish children and young teenager’s conversations: data preparation and short text classifcation with contextual word embeddings”.

[13]. F. Chollet, “Keras,” GitHub, 2015. Available: https://github.com/keras-team/keras.

[14]. J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, A. Mohammed, “Social media cyberbullying detection using machine learning,” IJACSA, 2019.