Protecting Children Online: Enhancing Hate Speech Detection on Twitter through AI Algorithms

Anıl Özfırat, Zafer Çınar, İsmail Deha Köse
ICT for Internet and Multimedia,
University of Padova, Padova, Italy
anil.oezfirat@studenti.unipd.it, zafer.cinar@studenti.unipd.it, ismaildeha.koese@studenti.unipd.it

Abstract

Hate speech is one of the most prevalent forms of toxic behavior on the Internet. One of the groups that are under the risk of exposure the most are children and adolescents. Since they are in the process of shaping their world view, being exposed to hateful content such as racism, misogyny or homophobia can cause psychological and social problems. Since the amount of data being transferred to the media is huge, we believe it is necessary to implement a solution that detects hate speech automatically. This study presents an approach to the problem by using the NLP method of the BERT model. We use data from Twitter to detect hate speech. Our study reports accuracy and F1 score of 0.91 and 0.78. We believe this result can be improved with more comprehensive data.

1. Introduction

With social media becoming one of the main means of communication, every person who is connected to the Internet is at risk of exposure to online harassment. Online harassment comes in many different forms such as cyberbullying which targets one individual or hate speech which targets a social group [1]. Despite these harmful content being all defined as online harassment, it is necessary to make a distinction between them to analyze their causes, effects and cultural aspects further. In the scope of this research, we decided to focus on hate speech which targets not an individual, but a social group or a minority.

Cambridge Dictionary defines hate speech as "public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation". Any person who is exposed to this hateful content is susceptible to psychological and social problems [2]. This can cause not only traumas and identity problems for the victims, but it can cause political radicalization and hate crimes on a social level [1]. One of the groups who are at the highest risk are children and adolescents. Delgado and Stefancic [1] argue that exposure to hate speech is especially dangerous for younger people because they are in the process of shaping their sense of identity and they usually lack the coping mechanisms that most adults do. This might make children feel a sense of inferiority, adapt a prejudicial world-view or sometimes more serious psychological problems that might remain throughout their life [3].

According to a report Facebook published [4], preventing hateful speech is considered as preservation of integrity and this principle is adopted by many social media platforms such as Twitter and Youtube. However, detecting and preventing hate speech presents a complex problem as it involves many different layers of cultural, historical and social aspects. Also, with respect to the principle of "free speech", the solutions should never fall within the boundaries of censorship.

To protect younger social media users from hate speech, we decided to develop a method for automatic detection using Natural Language Processing (NLP) method of BERT (Bidirectional Encoder Representations from Transformers). The problem we want to address is to be able process huge amounts of data in social networks and be able to be as precise as possible to eliminate unnecessary censorship.

2. Related Work

In the context of tweets, toxic language is essential to consider a broad range of harmful behaviors and their impact on various individuals and groups on social media platforms.[5,6,7,8] As defined by Reference [5], toxic tweets, encompassing elements such as harassment, threats, and offensive language targeted at specific groups. While our research especially focuses on hate speech, abusive language, cyberbullying, toxicity and trolling towards children, it is important to recognize that accounts engaging in hate speech often exhibit general toxic behavior towards not only ethnic, minorities but also LGBTQ communities and other social groups [5]. In this paper, distinguishing between hateful xenophobic content and general toxic online behavior, using the terms hateful and toxic respectively, while referring to the relationships between hate speech and related concepts discussed in Reference [9].

Davidson et al. [10] aim to distinguish the detection under three categories. It predicts corresponding these categories, If a text offensive, includes hate speech or is neutral. They are lowercasing each tweet and stemming tweets with using Porter stemmer, and then they create diagram, unigram, trigram features in texts. The best performing models of their approach have an overall precision, recall and F1 score of 0.91, 0.90 and 0.90. Mishra et al. [11] created two different graphs. The first one is identical to the community graph of the previous work, and the second graph is an extended version of the first. Previous research based on automated abusive language detection on Twitter. In short, working with graph convolutional networks to capture the problem, and analyzing heterogeneous graph information as part of the model. In order to collect the data and annotate the hate speech, they prepared a list of standards and analyzed the impact of various extra-linguistic features to enhance the quality of the standard [12,13].

In this paper, the data collected on German hate tweets following the 2017 German federal elections between August 2017 and April 2018. They provide an overview of right-wing hate speech from a legal and linguistic perspective on Twitter and segregated non-linguistic symbols, signs and words that express the ideology of users of hate speech [14].

As the use of the Internet by the younger generation accelerates, so has the research on the use of machine learning to protect children from hate speech. One recent work by Chiu et al. [15] aims at detecting hate speech with a specific focus on the users between age 13-15. This work utilizes three ML algorithms which are K- Nearest Neighbor (KNN), Naive Bayes and support vector machine (SVM) and aim at hate speech detection in collected 8100 tweets. The main idea of this research is that it uses TF-IDF (term frequency-inverse document frequency technique) to vectorize the data. After this data processing step, three models are trained. The first model is KNN model which is utilizing a k-value of 11. This model reports the precision, recall, and f1-score of 0.86, 0.88, and 0.88. The second model is Naive Bayes which gives the precision, recall, and f1-score of 0.95, 0.87, and 0.91. The third model is the SVM model which reports the precision, recall, and f1-score of 0.90, 0.98, and 0.94.
This shows that Naive Bayes is the best one with finding the true positives and supervised learning gives some decent results with detecting the text that includes hate speech.

Among the studies that investigate languages other than English, Polleto et al. [16] contribute valuable insights by discussing the intricate process and inherent challenges involved in curating a comprehensive hate speech corpus specifically tailored for Italian. Their work sheds light on the linguistic nuances and cultural context that shape hate speech in the Italian language, facilitating a deeper understanding of the phenomenon within this specific linguistic and cultural framework.

In a similar vein, Pereira et al. [17] focus their research on the Spanish language, tackling the pervasive issue of hate speech on Twitter. They develop HaterNet, an innovative system that not only identifies and classifies instances of hate speech but also provides an analysis of hate trends and other forms of negative sentiments within the Spanish-speaking Twitter community. By examining the patterns and dynamics of hate speech in Spanish, their work contributes to a broader understanding of the challenges faced in combating hate speech across different linguistic contexts and offers valuable insights for developing effective mitigation strategies.

Thanks to the continuous development and improvement of technology, we have access to numerous language models that can be utilized in this study. Among the most stable and up-to-date deep learning methods, we have chosen to employ BERT, a Transfer Learning-based natural language processing model[18,19]. BERT's bidirectional nature surpasses the unidirectional approach of GPT (Generative Pre-trained Transformer), and its working principle based on transfer learning enables it to outperform the LSTM learning method utilized by ELMO [20] in natural language processing. Additionally, BERT's ability to predict consecutive sentences allows us to analyze the entirety of a written tweet.

According to studies on text models, the Transfer Learning method is considered to be the most up-to-date method with the best results [18]. BERT has two novel prediction tasks: Masked LM (Language Model) and Next Sentence Prediction. This allows us to link consecutive sentences and analyze the whole text rather than the sentence [21]. To improve the performance of Bert's model and to deal with incivility in models is fine-tuned and also a CNN (Convolutional Neural Network) supported model will be created [22]. For this process it is proposed to use lexical features derived from a hate dictionary [23].

3. Dataset

The dataset used in this study is derived from Twitter data and was specifically collected to facilitate research on hate-speech detection. It encompasses a wide range of textual content, classified into three distinct categories: hate-speech, offensive language, and neither. It is crucial to emphasize that due to the nature of the study, the dataset contains text that may include instances of racism, sexism, homophobia, or content that is generally offensive.

The dataset used in this study is derived from Twitter data and was specifically collected to facilitate research on hate-speech detection. It encompasses a wide range of textual content, classified into three distinct categories: hate-speech, offensive language, and neither. It is crucial to emphasize that due to the nature of the study, the dataset contains text that may include instances of racism, sexism, homophobia, or content that is generally offensive.

4. Methodology

In this study, we adopted a text-centric approach, considering that the majority of the data used consisted of tweets. We developed a hate speech detection system based on Natural Language Processing (NLP) techniques using a curated dataset. The system was built by leveraging a language model to effectively identify instances of hate speech.

We utilized the BERT transformer model's performance in the task of hate speech detection. BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based neural network architecture that has achieved state-of-the-art results in various natural language processing tasks. The BERT model is pre-trained on a large corpus of unlabeled text data using two unsupervised learning tasks: masked language modeling (MLM) and next sentence prediction (NSP). MLM involves randomly masking some words in the input sentence and training the model to predict the masked words based on the context. NSP aims to determine whether two sentences appear consecutively in the original text or not. The pre-training process allows BERT to learn contextual representations of words and sentences [19].

In addition to BERT, we also utilized our model with Adam Optimizer for fine-tuning and a Convolutional Neural Network (CNN) architecture to capture local patterns and relationships within the text. The CNN model consisted of multiple convolutional layers followed by non-linear activation functions. Instead of solely relying on the output of the final transformer encoder, we employed the outputs of all transformer encoders. This integration allowed us to leverage the complementary strengths of both BERT and CNN in capturing intricate textual features.[22].

4.1. Text Modality

We trained our model by using the BERT. We used two types of BERT, L=12 hidden layers (i.e., Transformer blocks), a hidden size of H=768 with 5 epochs, and A=12 attention heads and L=4, H=512, A=4 with 20 epochs. The tweets which in our dataset have been lower-cased and labeled before they were fed to the classifier. We used Sparse Cross Entropy Loss and fine-tuned the network using Adam Optimiser with initial learning rate init_lr = 3x10−5. Detailed result shown as figures and tables below:

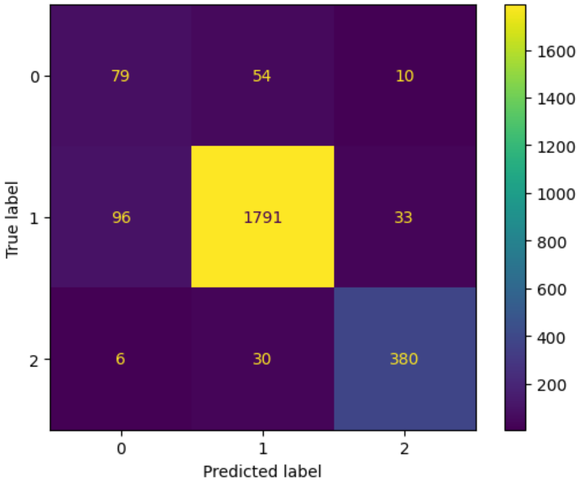

Figure 1 BERT L=12, H=768, A=12 Confusion Matrix

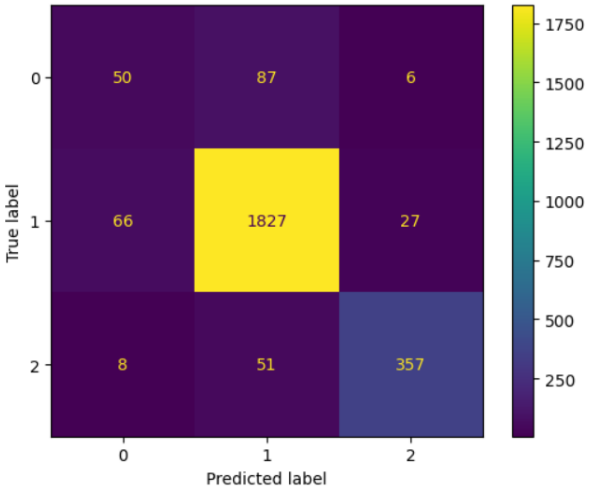

Figure 2 BERT L=4, H=512, A=4 Confusion Matrix

Table 1 Detailed Results

Model	*Accuracy*	*F1-Score (Macro)*
BERT L=12, H=768, A=12	0.91	0.78
BERT L=4, H=512, A=4	0.90	0.73

5. Results and Conclusions

Upon analyzing the results obtained from two different BERT models, we observed that BERT L=12, H=768, A=12 produced more accurate results. The reason for this can be attributed to the fact that BERT L=12, H=768, A=12 has a more advanced and deeper word embedding due to its neural network architecture. As evident from the confusion matrices, there were challenges in distinguishing hate speech and offensive speech by the model. Furthermore, the model successfully distinguished tweets containing hate speech and offensive language from others.

In this paper which observed twitter hate speech dataset to protect children, despite the high accuracy rate, it was observed that hate speech can be confused with offensive speech. This problem can be overcome with more data and more training. As a different solution, image processing can be added to the text approach and clearer results can be obtained through both visual and written analysis. Offensive speech can have a negative impact on their perspective and personality as hate speech can have a negative impact on children. However, the results obtained are capable of protecting children from twitter.

6. Future Work & Thoughts

The effect of children on social media hate speech is a pressing concern that necessitates comprehensive solutions. One crucial step is the development of algorithms capable of working simultaneously in multiple known languages. These algorithms should be implemented by all social media platforms to foster positive child development. By integrating such algorithms, platforms can effectively combat social harassment, cyberbullying, scams, and other harmful activities. Creating an online environment that prioritizes child safety and well-being is paramount. With the addition of these applications or algorithms, social media platforms can significantly reduce the prevalence of hate speech and ensure a healthier online experience for children and users of all ages.

References

1. Richard Delgado and Jean Stefancic, “Understanding Words that Wound”, 2004, Westview Press
2. Krause, N., Ballaschk, C., Schulze-Reichelt, F., Kansok-Dusche, J., Wachs, S., Schubarth, W., & Bilz, L. (2021). “Ich lass mich da nicht klein machen!“ – Eine qualitative Studie zur Bewältigung ¨ von Hate Speech durch Schüler/innen [I don’t let them get me down!”—A qualitative study on students’ coping with hate speech]”. Zeitschrift für Bildungsforschung, 11(1), 169–185.
3. Simpson, Robert. (2019). ‘Won’t Somebody Please Think of the Children?’ Hate Speech, Harm, and Childhood. Law and Philosophy. 38. 10.1007/s10982-018-9339-3.
4. Halevy, A.; Ferrer, C.C.; Ma, H.; Ozertem, U.; Pantel, P.; Saeidi, M.; Silvestri, F.; Stoyanov, V. Preserving Integrity in Online Social Networks. arXiv 2020, arXiv:2009.10311
5. Radfar, B.; Shivaram, K.; Culotta, A. Characterizing Variation in Toxic Language by Social Context. In Proceedings of the International AAAI Conference on Web and Social Media; Association for the Advancement of Artificial Intelligence: Menlo Park, CA, USA, 2020; Volume 14, pp. 959–963.
6. Guberman, J.; Schmitz, C.; Hemphill, L. Quantifying toxicity and verbal violence on Twitter. In Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion, San Francisco, CA, USA, 27 February–2 March 2016; pp. 277–280.
7. Gunasekara, I.; Nejadgholi, I. A review of standard text classification practices for multi-label toxicity identification of online content. In Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Brussels, Belgium, 31 October–1 November 2018; pp. 21–25.
8. Parent, M.C.; Gobble, T.D.; Rochlen, A. Social media behavior, toxic masculinity, and depression. Psychol. Men Masculinities 2019, 20, 277.
9. Poletto, F.; Basile, V.; Sanguinetti, M.; Bosco, C.; Patti, V. Resources and benchmark corpora for hate speech detection: A systematic review. Lang. Resour. Eval. 2020, 55, 477–523 .
10. Davidson, T.;Warmsley, D.; Macy, M.;Weber, I. Automated Hate Speech Detection and the Problem of Offensive Language. arXiv 2017, arXiv:1703.04009.

11. Mishra, P.; Del Tredici, M.; Yannakoudakis, H.; Shutova, E. Abusive Language Detection with Graph Convolutional Networks. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 2145–2150.
12. Waseem, Z.; Hovy, D. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In Proceedings of the NAACL Student Research Workshop, San Diego, CA, USA, 13–15 June 2016; Association for Computational Linguistics: San Diego, CA, USA, 2016; pp. 88–93.
13. Waseem, Z. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In Proceedings of the FirstWorkshop on NLP and Computational Social Science, Austin, TX, USA, 5 November 2016; Association for Computational Linguistics: Austin, TX, USA, 2016; pp. 138–142.
14. Jaki, S.; Smedt, T.D. Right-wing German Hate Speech on Twitter: Analysis and Automatic Detection. arXiv 2019, arXiv:1910.07518
15. A. Chiu, K. Sood, A. Rincon and D. Doran, "Detecting Hate Speech on Social Media with Respect to Adolescent Vulnerability," 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2023, pp. 0724-0728, doi: 10.1109/CCWC57344.2023.10099373.
16. Poletto, F.; Stranisci, M.; Sanguinetti, M.; Patti, V.; Bosco, C. Hate speech annotation: Analysis of an italian twitter corpus. In Proceedings of the 4th Italian Conference on Computational Linguistics, CLiC-it 2017, CEUR-WS, Rome, Italy, 11–13 December 2017; Volume 2006, pp. 1–6.
17. Pereira-Kohatsu, J.C.; Quijano-Sánchez, L.; Liberatore, F.; Camacho-Collados, M. Detecting and monitoring hate speech in Twitter. Sensors 2019, 19, 4654.
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In Advances in Neural Information Processing Systems; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S.,Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2017 ; Volume 30.
19. Devlin, J., Chang, M.W., Lee, K, Toutanova, K., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, 2019, Proceedings of NAACL-HLT 2019, MN, USA, pp-4171-4186
20. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L. “Deep contextualized word representations”, 2018, NAACL 2018, arXiv:1802.05365v2
21. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., “Improving Language Understanding by Generative Pre-Training”
22. Ozler, K.B.; Kenski, K.; Rains, S.; Shmargad, Y.; Coe, K.; Bethard, S. Fine-tuning for multi-domain and multi-label uncivil language detection. In Proceedings of the Fourth Workshop on Online Abuse and Harms, 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 28–33.
23. Koufakou, A.; Pamungkas, E.W.; Basile, V.; Patti, V. HurtBERT: Incorporating Lexical Features with BERT for the Detection of Abusive Language. In Proceedings of the Fourth Workshop on Online Abuse and Harms, Online, 20 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 34–43.