Hate Speech and Offensive Content Identification in English Language Based on BERT Model

Jing Li, Xutao Yang*

School of information Science and Engineering, Yunnan University, Kunming, 650500, Yunnan, P.R. China

Abstract

With the development of the Internet, people are becoming increasingly interconnected through the network. More and more people interact through social media such as Weibo, YouTube, Facebook, etc. While this kind of interaction makes people's connection closer, it also brings problems. Some individuals post hate and offensive language online to attack others, which not only damages the good online atmosphere but also harms the physical and mental health of the victims. Therefore, it is crucial to prohibit the occurrence of hate and offensive language on these social media platforms. Based on the BERT model, this paper explores three ways to detect hate speech and offensive content in English at FIRE 2024 task-1. The BERT and Convolutional Neural Network (CNN) method achieves macro F1 scores of 0.80049 on the HASOC 2024 English test set. The BERT and Recurrent Neural Network (RNN) method achieves macro F1 scores of 0.79075. For the improved Recurrent Neural network, it achieves macro F1 scores of 0.78065 on the same test set. Among the three methods, the BERT and Convolutional Neural Network combination model achieves the highest score, performs the best, and ranks 2nd in Task-1.

Keywords

hate speech, offensive content, English, BERT, HASOC 2024

1. Introduction

With the development of the Internet era, more and more people are using social media platforms. Through this medium, people worldwide can chat, make friends, share experiences, express opinions, and even brainstorm. These are the positive aspects that social media platforms bring to us. However, social media platforms also have negative impacts. They often contain hate and offensive content, such as pejorative[1], cyber-bullying [2, 3], online extremism [4], and racism [5, 6]. Such content may cause anxiety in those who are targeted, leading to psychological issues and subjecting others to violence[7]. It has been recognized that online hate speech is a social issue that harms our society[8]. It is essential to detect such hate and offensive content early and prevent its spread.

Since 2019, HASOC has been dedicated to sharing tasks for the detection of hate speech and offensive content, as well as researching languages, including English and some low-resource languages such as Marathi. In 2024, HASOC shared this task: Hate Speech and Offensive Content Identification in English and Bangla[9, 10]. The task has two sub-tasks, and we participated in Task 1.

The main task of Task 1 is to classify the given English Twitter content into two classes:

- Hate and Offensive (HOF): post contains hate, offensive, and profane content.
- None Hate-Offensive (NOT): post contains no Hate speech, profane, offensive content.

For this task, we use the BERT model to encode English sentences. Then we use three types of networks for further classification: Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and an improved version of Recurrent Neural Network. Ultimately, among these three methods, the combination of the BERT model with Convolutional Neural Network performs the best, achieving a macro F1 score of 0.80049, while RNN and an improved version of RNN have macro F1 scores of 0.79075 and 0.78065 respectively.

Section 2 introduces the methods used in recent years to detect hate and offensive content in English. Section 3 describes the source of the dataset used in this task, the division of the dataset, and the test

Forum for Information Retrieval Evaluation, December 12-15, 2024, India

*Corresponding author.

jing61167@gmail.com (J. Li); yangxutao@ynu.edu.cn (X. Yang)

set. In Section 4, we carefully describe the dataset's preprocessing, the model's workflow, and some parameter settings during the model training process. Section 5 presents the macro F1 scores achieved by the three methods on the HASOC 2024 test set. In the final section, we summarize this task's work, point out our experiment's shortcomings, and look forward to the future.

2. Related Work

The detection of hate and offensive content has been researched for a long time. HASOC has been dedicated to detecting hate and offensive content from 2019 to 2024, involving languages including English and some low-resource languages such as Marathi. An increasing number of papers are getting engaged in detecting hate and offensive content. With their efforts, the classification performance of models detecting hate and offensive content has been continuously improving.

Typically, the detection of hate and offensive content is approached as a supervised classification task[11], such as a binary classification task, to determine whether a sentence belongs to hate and offensive content. In the early days, some researchers used lexicon-based feature methods to detect hate and offensive content[12]. However, their proposed models could not distinguish between hate content and offensive content. At present, the machine learning methods are widely used. Traditional machine learning methods often require manual feature engineering to extract text features, while deep learning-based methods mainly use neural networks. Risch and Krestel[13] used the traditional machine learning method, logistic regression classifier, to detect offensive content. Waseem[14] used SVM and logistic regression classifier to detect racist or sexist content. In 2018, Pitsilis et al.[15] used deep learning models, such as Recurrent Neural Networks, to detect offensive content in English. Recurrent Neural Networks consider the inputs from previous time steps when processing the current input, which helps them understand contextual information. So, it achieves good results in the field of natural language. As a variant of recurrent neural networks, long short-term memory networks have also achieved good results in detecting hate and offensive content[16]. Since 2020, language models based on transformers, such as the BERT, m-BERT, and XLM-RoBERTa models[17], have become increasingly popular in classification tasks. In the HASOC 2020 task, the YNU_OXZ team[18] proposed a model based on XLM-RoBERTa and LSTMs for hate speech detection in English. In the HASOC 2021, the team using a Graph Convolutional Network approach achieved the best results in detecting hate and offensive content in English[19].

Due to the excellent sentence representation capability of transformer-based models and the ability of Convolutional Neural Networks to capture local features, it may be able to detect insulting words within sentences. So, we use the combination of the BERT model and CNN to detect hate and offensive content. At the same time, because a RNN can detect hate and offensive content effectively, so we also use the BERT model to encode English sentences and feed the last_hidden_state of the model into the RNN for classification. In this way, we compare the results of these three methods to obtain a more effective approach.

3. Datasets

This task focuses on the binary classification of hate and offensive speech in English. The organizers didn't offer the training data for the datasets but provided the test data collected from Twitter. Therefore, we downloaded the English datasets from HASOC 2019[20] and HASOC 2020[21], which have 5852 and 3792 tweets, respectively. Then, we divided the validation set and training set in a ratio of 2:8. So, we have 7715 training data in total, 1929 validation data, and 885 test data. The data are all stored in CSV file format. The training set includes sentence id, text content, and labels (HOF or NOT), while the test set is without labels. Table 1 shows examples of hate and offensive content in the training set. In Figure 1, we have presented the data distribution in the training and validation set for binary classification.

Table 1
the Examples of the Hate and Offensive Content in Training Set

Id	Comments	Label
hasoc_en_1	#DhoniKeepsTheGlove | WATCH: Sports Minister Kiren Rijiju issues statement backing MS Dhoni over 'Balidaan Badge', tells BCCI to take up the matter with ICC and keep government in the know as nation's pride is involved https://t.co/zuo5335Rjr.	NOT
hasoc_en_2	@politico No. We should remember very clearly that #Individual1 just admitted to treason . #TrumpIsATraitor#McCainsAHero #JohnMcCainDay"	HOF

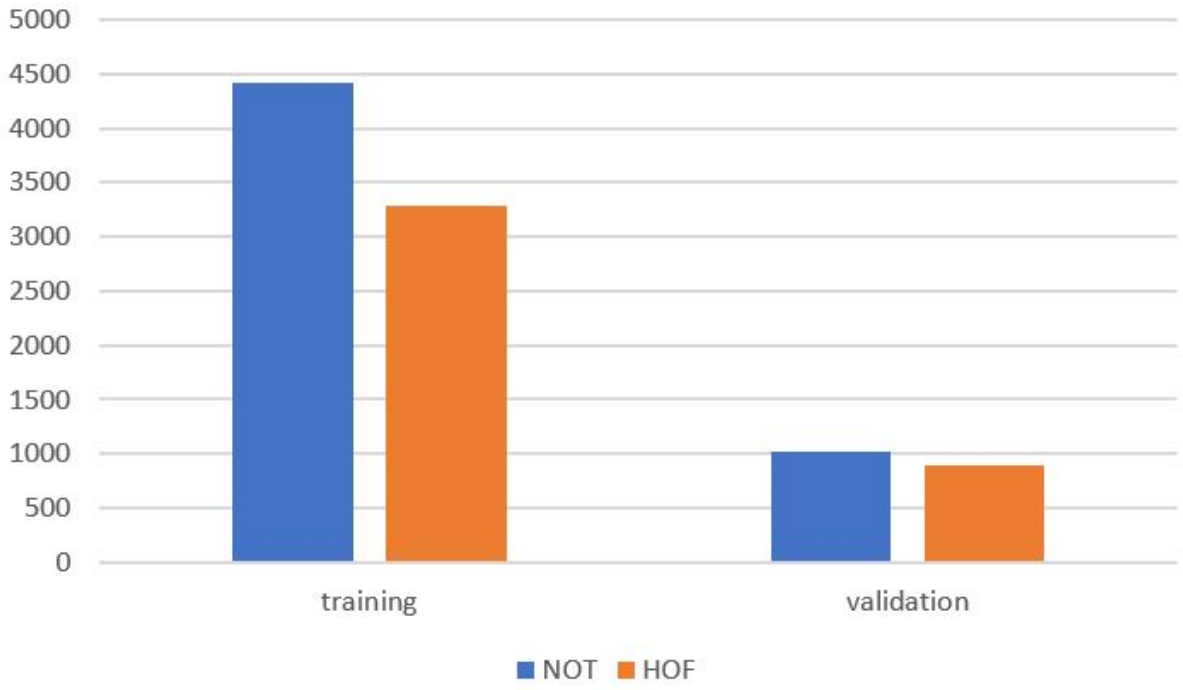

Figure 1: Data Distribution in Training and Validation Set.

4. Experiment

In our experiment, we first preprocess the datasets to remove special characters from English sentences. Then, we use three methods based on the BERT model for classification. We use the BERT model to represent the vector representation of English sentences and pass the vector representation of the sentences through three types of networks for classification: Convolutional Neural Network, Recurrent Neural Network, and an improved version of Recurrent Neural Network. We search for the most effective model among them. Below, we describe our experiment in detail.

4.1. Preprocessing

After reviewing the datasets, we find many special characters in English sentences, such as emoji emoticons, URLs, mentions (@someone), and numbers. These special characters can affect the encoding effectiveness of the model, so we take some measures to handle these special characters. Our steps are as follows:

- Replace all numbers with the word "number".
- Replace emojis with corresponding text descriptions because some emotion may be included in emojis.

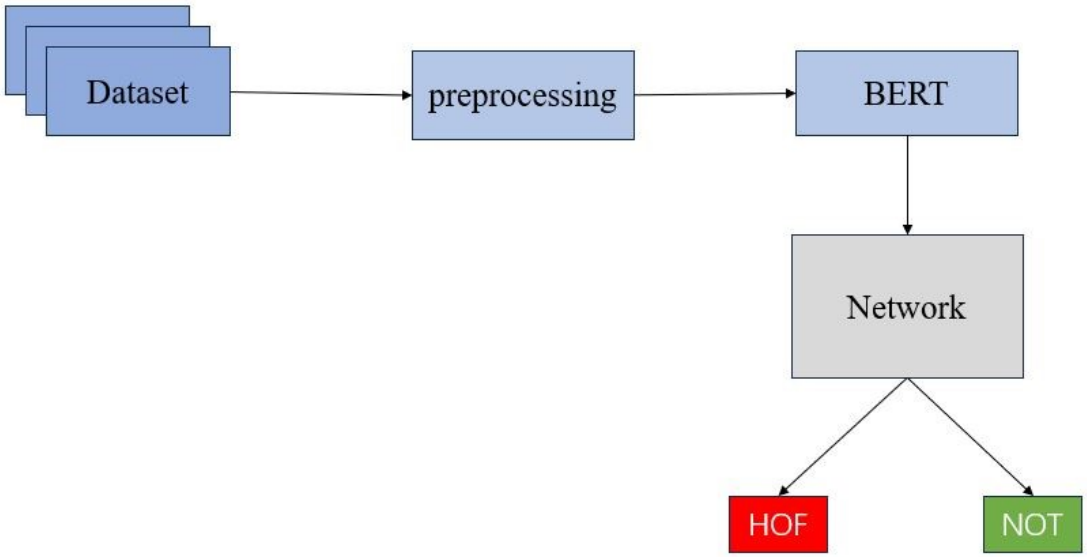

Figure 2: Overall Structure of the Model.

- Replace all website links with "URL".
- Remove all the tagged user name.
- Remove all the hashtags.
- Remove all punctuation marks.

We have also processed the labels. We replace HOF with 1 and NOT with 0 in the labels, which makes it more convenient for training classification tasks.

4.2. Modeling

As we all know, the BERT model is based on a bidirectional transformer structure, which can capture the contextual information on both sides of English words and understand complex semantic relationships effectively. Therefore, we use the BERT model for the vector representation of English sentences, and then send the vector representation to three types of networks for further classification. Figure 2 is the overall structure of the model using our methods.

The first method involves feeding the sentence embedding represented by the BERT model into a Convolutional Neural Network. First, we expand the dimension of the sentence vectors obtained from BERT, and then input them into a two-dimensional convolutional neural network. The vector is passed sequentially through convolutional layers, pooling layers, and finally fully connected feed-forward network layer. In this method, we use convolutional kernels of different sizes to increase the diversity of feature extraction. The kernel sizes we use are 2×768, 3×768, 4×768.

The second method involves feeding the sentence embedding represented by the BERT model into a Recurrent Neural Network. We implement the RNN using Long Short-Term Memory (LSTM) networks. We take the hidden state of the last time step of the LSTM layer and pass it into a fully connected feed-forward network layer.

The third method is based on the second method. We modify the second method slightly. We perform a max-pooling operation on the output of the LSTM layer, which can extract the maximum values along the last dimension, and then the result is passed into a fully connected feed-forward network layer.

Table 2
Macro F1 Scores of Our Three Methods

Model	Macro F1 Score
BERT+CNN	0.80049
BERT+RNN	0.79075
BERT+ improved RNN	0.78065

4.3. Training

During the model training process, we use the training set to train the model and the validation set to find the optimal model parameters, which are then applied to the test set for prediction. Since this task is a binary classification task, we use the cross-entropy function to calculate the loss and macro F1 score as the evaluation metric to assess the model's classification performance. Additionally, since the amount of training data is insufficient, we employ early stopping and learning rate scheduling strategies to prevent the model from overfitting. We use the training sets of HASOC 2019 and HASOC 2020 as the dataset and train on three methods. The hyperparameter values and number of epochs remain the same across all three methods in our experiment. Ultimately, we find that the combination of the BERT model and CNN method performs best on the test set of HASOC 2024. This method uses AdamW as the optimizer with a learning rate of 1e-5 and a batch size of 32.

5. Results

The competition for Task 1 is evaluated on macro-f1 metrics. There are a total of 8 teams participating in Task 1. And our team ranks 2nd in Task-1. The final results are shown in Table 2.

Table 2 shows that the combination of the BERT model and CNN performs slightly better in the binary classification of hate and offensive content in English, which gets the macro F1 score of 0.80049.

6. Conclusion

This paper describes our methods for detecting hate and offensive content in English. We conduct experiments using three BERT-based methods, in which the BERT and CNN combination model performs best. It may be because CNN captures local features effectively, which helps the model capture the specific patterns of hate or offensive language in sentences. So, the method based on BERT and CNN performs a little better than the method based on BERT and RNN. In the future, we plan to research the detection of hate and offensive content in low-resource languages. At the same time, due to the limited data resources available for low-resource languages, we may explore research in few-shot learning. We hope that in the future, we can have a clean and righteous online space.

Acknowledgments

Thanks to the organizers of the competition. And this work is supported by Scientific Research and Innovation Project of Postgraduates Students in the Academic Degree of YunNan University (KC-242410495).

Declaration on Generative AI

We have not employed any Generative AI tools.

References

[1] L. P. Dinu, I. Iordache, A. S. Uban, M. Zampieri, A computational exploration of pejorative language in social media, in: M. Moens, X. Huang, L. Specia, S. W. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021, Association for Computational Linguistics, 2021, pp. 3493–3498. doi:10.18653/V1/2021.FINDINGS-EMNLP.296.

[2] H. Rosa, N. S. Pereira, R. Ribeiro, P. C. Ferreira, J. P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A. M. V. Simão, I. Trancoso, Automatic cyberbullying detection: A systematic review, Computers in Human Behavior 93 (2019) 333–345. doi:10.1016/J.CHB.2018.12.021.

[3] J. Shetty, K. N. Chaithali, A. M. Shetty, B. Varsha, V. Puthran, Cyber-bullying detection: A comparative analysis of twitter data, in: N. N. Chiplunkar, T. Fukao (Eds.), Advances in Artificial Intelligence and Data Engineering, Springer Singapore, Singapore, 2021, pp. 841–855.

[4] S. Aldera, A. Z. Emam, M. Al-Qurishi, M. A. AlRubaian, A. Alothaim, Online extremism detection in textual content: A systematic literature review, IEEE Access 9 (2021) 42384–42396. doi:10.1109/ACCESS.2021.3064178.

[5] I. Kwok, Y. Wang, Locate the hate: Detecting tweets against blacks, in: M. desJardins, M. L. Littman (Eds.), Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, July 14-18, 2013, Bellevue, Washington, USA, AAAI Press, 2013, pp. 1621–1622. doi:10.1609/AAAI.V27I1.8539.

[6] S. Tulkens, L. Hilte, E. Lodewyckx, B. Verhoeven, W. Daelemans, A dictionary-based approach to racism detection in dutch social media, CoRR abs/1608.08738 (2016). URL: http://arxiv.org/abs/1608.08738. arXiv:1608.08738.

[7] R. Bannink, S. Broeren, P. M. van de Looij-Jansen, F. G. de Waart, H. Raat, Cyber and traditional bullying victimization as a risk factor for mental health problems and suicidal ideation in adolescents, PLoS ONE 9 (2014).

[8] M. L. Williams, P. Burnap, A. Javed, H. Liu, S. Ozalp, Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime, The British Journal of Criminology 60 (2019).

[9] N. Raihan, K. Ghosh, S. Modha, S. Satapara, T. Gaur, Y. Dave, M. Zampieri, S. Jaki, T. Mandl, Overview of the hasoc track at fire 2024: Hate-speech identification in english and bengali, in: Forum for Information Retrieval Evaluation (FIRE 2024) Working Notes, CEUR-WS.org, Gandhinagar, India, 2024. December 9–13, Gandhinagar, India.

[10] K. Ghosh, N. Raihan, S. Modha, S. Satapara, T. Gaur, Y. Dave, M. Zampieri, S. Jaki, T. Mandl, Overview of the hasoc track at fire 2024: Hate-speech identification in english and bengali, in: Proceedings of the 16th Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE '24), Association for Computing Machinery, New York, NY, USA, 2024.

[11] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: L. Ku, C. Li (Eds.), Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, SocialNLP@EACL 2017, Valencia, Spain, April 3, 2017, Association for Computational Linguistics, 2017, pp. 1–10. doi:10.18653/V1/W17-1101.

[12] Y. Chen, Y. Zhou, S. Zhu, H. Xu, Detecting offensive language in social media to protect adolescent online safety, in: 2012 International Conference on Privacy, Security, Risk and Trust, PASSAT 2012, and 2012 International Confernece on Social Computing, SocialCom 2012, Amsterdam, Netherlands, September 3-5, 2012, IEEE Computer Society, 2012, pp. 71–80. doi:10.1109/SOCIALCOM-PASSAT.2012.55.

[13] C. Nobata, J. R. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive language detection in online user content, in: J. Bourdeau, J. Hendler, R. Nkambou, I. Horrocks, B. Y. Zhao (Eds.), Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016, ACM, 2016, pp. 145–153. doi:10.1145/2872427.2883062.

[14] Z. Waseem, Are you a racist or am I seeing things? annotator influence on hate speech detection on twitter, in: D. Bamman, A. S. Dogruöz, J. Eisenstein, D. Hovy, D. Jurgens, B. O'Connor,

A. Oh, O. Tsur, S. Volkova (Eds.), Proceedings of the First Workshop on NLP and Computational Social Science, NLP+CSS@EMNLP 2016, Austin, TX, USA, November 5, 2016, Association for Computational Linguistics, 2016, pp. 138–142. doi:10.18653/V1/W16-5618.

[15] G. K. Pitsilis, H. Ramampiaro, H. Langseth, Detecting offensive language in tweets using deep learning, CoRR abs/1801.04433 (2018). URL: http://arxiv.org/abs/1801.04433. arXiv:1801.04433.

[16] G. L. D. la Peña Sarracén, R. G. Pons, C. E. Muñiz-Cuza, P. Rosso, Hate speech detection using attention-based LSTM, in: T. Caselli, N. Novielli, V. Patti, P. Rosso (Eds.), Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018, volume 2263 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018. URL: https://ceur-ws.org/Vol-2263/paper040.pdf.

[17] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. doi:10.18653/V1/N19-1423.

[18] X. Ou, H. Li, Ynu_oxz at hasoc 2020: Multilingual hate speech and offensive content identification based on xlm-roberta, in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, Hyderabad, India, 2020, pp. 121–127. URL: http://ceur-ws.org/Vol-2826/T2-3.pdf, december 16–20, 2020.

[19] T. Mandl, S. Modha, G. K. Shahi, et al., Overview of the hasoc subtrack at fire 2021: Hatespeech and offensive content identification in english and indo-aryan languages, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, volume 3159 of *CEUR Workshop Proceedings*, CEUR-WS.org, Gandhinagar, India, 2021, pp. 1–19. URL: https://ceur-ws.org/Vol-3159/T1-1.pdf, december 13–17, 2021.

[20] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandalia, A. Patel, Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages, in: FIRE '19: Forum for Information Retrieval Evaluation, Association for Computing Machinery, Kolkata, India, 2019, pp. 14–17. URL: https://doi.org/10.1145/3368567.3368584. doi:10.1145/3368567.3368584, december 2019.

[21] T. Mandl, S. Modha, G. K. Shahi, A. K. Jaiswal, D. Nandini, D. Patel, P. Majumder, J. Schäfer, Overview of the HASOC track at FIRE 2020: Hate speech and offensive content identification in indo-european languages, CoRR abs/2108.05927 (2021). URL: https://arxiv.org/abs/2108.05927. arXiv:2108.05927.