

Advanced Deep Learning Approaches for Hate Speech Detection

Qizheng Wang

College of Computer Science and Cyber Security (Oxford Brooks College), Chengdu University of Technology, Chengdu, 610059, China

herewang.qizheng@zy.cdut.edu.cn

Abstract. At present, the influence of social networks on a global scale is increasing, the number of users is increasing significantly, and massive amounts of information are being created every day. How to automatically and efficiently identify harmful information, especially hate speech, has become an important issue in the governance of the network environment. In order to improve the screening efficiency of information containing hate speech, deep learning is applied in this field. Although the research work around this task has made great progress, there are very few reviews on this task, lack of a comprehensive review of the latest development in recent years and can not provide help for researchers who are interested in this task. Therefore, we give an overview of the deep learning applied in hate speech, introducing new ideas for solving this task in recent years, and propose potential problems in the task and analyze it.

Keywords: Deep Learning, detection model, language analysis model, classified dataset.

1. Introduction

With the continuous rise of social media, people can express their opinions anytime and anywhere, but now social media platforms are gradually becoming platforms for spreading hate, even violence, threatening speech, and even a trend of flooding. This is not unfounded. According to an EU survey, 80% of social media users said they had faced online hate speech; 40% said they had been attacked and threatened online [1]. In many social media, such as Twitter, Facebook, etc., have tried to screen these speeches manually, but the efficiency is not high.

Previously, in order to manage hate speech automatically, researchers have also tried to classify content through traditional machine learning methods, such as using SVM-based kernel method or using Bayesian network to build models [4]. However, traditional machine learning is based on feature engineering and relies on It is used to manually construct data features, and then extract features from the data for machine learning algorithms or models to classify or regress tasks. Traditional machine learning is based on feature engineering, which relies on artificially constructing data features, and then extracting features from the data for machine learning algorithms or models to classify or regress tasks. there are two problems with this type of method: firstly, the construction of task features depends on human background knowledge, so the quality of the features is unstable; secondly, these features are shallow, mostly based on statistical methods, and cannot be very good. It also means that models obtained using such methods cannot identify hate speech with complex content and meaning. Different from the previous practice of obtaining data features, data (including text, images, etc.) are directly processed as real-valued vectors for learning [2]. Deep learning proposes the concept of end-to-end training, which saves the data feature annotation before each independent learning task is executed, which fully shows the representability of the deep learning model. Also, deep learning models could learn task-related semantic information automatically through multiple layers of nonlinear transformations, for example, in models for translation, the model can be built directly from the corpus without human intervention [2]. Using similar properties, some researchers in recent years have applied deep learning to the task of hate speech detection.

Although deep neural networks have been applied to hate speech detection, many problems have been solved and the performance in this area has been improved. However, there are very few reviews on this aspect so far. As a result, many researchers who want to start the task of hate speech detection

do not have a good understanding of the latest progress in the field and what new methods have been proposed. Therefore, we have sorted out and summarized the development of this task in recent years. We first differentiate the definitions of hate speech and specifically explain the necessity of using deep learning models, then introduce the commonly used datasets and metrics for the task. In addition, we also explain the mainstream deep learning methods used in hate speech detection models. Finally, we introduce some state-of-the-art methods in tasks and potential challenges in tasks.

2. Background

2.1. Overview of hate speech

2.1.1 What is hate speech?

As mentioned above, social media platforms have now become the hardest hit areas for hate speech, and social media platforms have also defined hate speech in their Community Standards and Policy; it would also be used for managing the environment. At the same time, the creators of the dataset would follow the similarly standards to dividing the text.

Facebook is the most used social media platform in the world and it defines the hate speech as a direct attack against people on the basis of their race, ethnicity, national origin, disability, sexual orientation, sex, gender identity and serious disease religious affiliation, caste etc. The method contains dehumanising speech, harmful stereotypes, statements of inferiority, cursing and calls for exclusion or segregation, expressions of contempt, disgust or dismissal [12]. At present, Facebook has started to use machine learning and related algorithms to check hate speech manually, and posts detected as hate speech will be deleted.

As the second most used social media in the world, twitter defines hate speech as against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, disability, serious disease, sexual orientation, gender, gender identity, religious affiliation, age, etc. Twitter will block accounts with direct violent speech, and comments with hate speech will be made unreadable [11]. What's more, in some countries, the hate speech is described in the laws as the speeches inciting violence, intimidating, disparaging, prejudicial actions against a group, or individuals, on the basis of the membership in a group [15].

2.1.2 What are the categories of hate speech?

In the process of managing hate speech, we should learn as much as possible about the types of hate speech, otherwise the trained model may only recognize some existing hate speech. The main types of hate speech include: General hate speech, Racism, Sexism, Religion Anti-semitism, Nationality, Other Physical/mental handicap, Politics, Sectarianism Socio-economical status [5].

2.1.3 What kind of text is judged as hate speech

Before discussing what kind of content is hate speech, it must be noted that in most of the data sets, whether it is hate speech is manually marked. At present, most of the detection tasks are also completed by machine learning models and manual Cooperation (the complaints that have been misjudged are completely handled manually). Therefore, whether a speech is "hate speech" inevitably has some subjectivity.

However, this should also be considered in light of specific circumstances. For example, the algorithm review agency of Facebook once judged the content in the American Declaration of independence as hate speech. If a certain group of people is describing their own experiences, for example, "mother with three children" is complaining about gender discrimination, although the text may have the same context and use discriminatory words, it does not belong to hate speech [9]. If the model for screening hate speech does not recognize this, the voice of these vulnerable groups will be impaired. Nowadays, various social platforms are popular with "humor" with some hatred, and some platforms (such as Facebook) think it is reasonable, which makes it more difficult to judge whether a word is racial discrimination.

To sum up, the current speech environment is becoming more and more complex, and the preliminary algorithm judgment and machine learning can not meet the requirements. Therefore, deep learning and neural network should be introduced to enhance the performance of the model to solve such problems.

2.2. Datasets

Different datasets are used in the training process of different models. Here we will introduce the most used and latest datasets. The Waseem dataset contains 16,914 tweets from Twitter, which is one of the most important social medias in the world. The amount of text is limited to 140 words on Twitter, which is convenient for text analysis. And the Similarly, in 2018, AGARWAL created the Twitter Hate Speech dataset on Kaggle, which is also widely used for training the model. Davidson collected 24783 tweets in 2017, dividing them into offensive speech, hate speech, and neither.

In addition to Twitter, there are also many datasets that gather information from different media: Djuric et al collect the information from Yahoo news and Yahoo finance in 2015, and Wulczyn collected 115,737 instances from Wikipedia where 88% of them is legitimate speeches. What's more, Degibert collected hate and legitimate speeches in 2018 in a white supremacy forum 'Stormfont. [3]

2.3. Metrics

According to the current research on hate speech detection, two commonly used indicators, Accuracy and F1 score, are very effective in judging the accuracy of the model.

2.3.1 Method for calculating Accuracy

The model would be applied to the test dataset, and the result would be divided into: TP (The actual result and prediction result is positive), FP (The actual result is positive but the prediction result is negative), FN (The actual result is false but the prediction result is positive), TN (The actual result and prediction is negative). The accuracy would be calculated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

2.3.2 Method for calculating F1 score

Before calculating F1 score, the Precision and Recall of the model would be calculated. The Precision is the proportion of positive classes in the samples identified as positive classes:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Similarly, call is the proportion of positive classes in the samples identified as positive classes:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

The F1-Measure is calculated as the weighted harmonic average of precision and recall:

$$F = \frac{(a^2+1)*P*R}{a^2*(P+R)} \quad (4)$$

In the formula, a is the constant that needs to be specified, when a equal to 1, the formula would turn to:

$$F1 = \frac{2*P*R}{P+R} \quad (5)$$

F1 combines the precision and recall. The higher the value of F1 score shows the model's abilities.

3. Recent Researches

In this part, some of the latest ideas and methods used to build and improve the model will be introduced, and the test results of the researchers will be provided for reference and inspiration for future researchers.

3.1. HABERTOR

The Model HABERTOR is developed from the BERT (Bidirectional Encoder Representations from Transformers) model, which is used for natural language analyzing. Comparing with the original BERT model from google, words from data for HABERTOR are embedded by Quaternion Factorization and compression technology, which make the sentence representation has less parameters, the original 110M data only uses 8M parameters. It gives the HABERTOR model faster training and inferencing, less memory usage, and a better accuracy. The model is inherited from the Bert model, which means the model has similar architecture and training progress.

It needs to be pre-trained to obtain basic semantic parameters and then used for specific tasks. This step replaces manual labeling of the text (but the original data is already labelled). In this step, the training dataset would be generated. In the training data generated by the original Bert model, 80% of the words in the sentences are replaced by [mask]. The improved training process of this model uses different alternatives: set τ as a parameter, use the replacement mask position τ times, and randomly sample to generate τ training examples to improve the model training ability. After that the HABERTOR made a Next sentence prediction task for improving the abilities of capturing the connection of two sentences, so that the model could understand the connection of long sequence context. According to the original BERT model, the pre-trained model will be further trained for specific scenarios, in this situation, the model will be used to classify the text as hate speech or not.

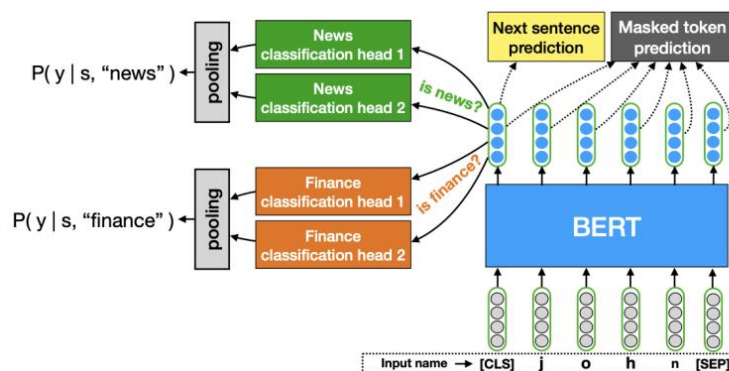


Figure 1. Model for HABERTOR

This simulates a real situation, the semantics of different sources may be different, which further improves the effectiveness of the model. The dataset used for testing the model explores on Yahoo News and Yahoo Finance and collected 1.4M comments extracted from, classified as: "is hate speech" or "not hate speech". The bottom shows the different improvement for the dataset, combine the model with different factorization, such as Vocab, Attention, Feedforward, Output Quaternion Factorization.

3.2. Detecting Hate Speeches with User Modeling

To improve the ability of classifying hate speech, not only the meaning of the text will be considered, but the status of the person making the comment is also used as a parameter to measure whether the speech is hate speech. These features of users, also called 'metadata' are used for building the model: Gender (male and female), Network (number of followers), Activity (number of tweets and the favorite) and the status of the profile, including: Geo (enable or not), default profile (yes or no), default image (yes or no), number of public lists [9]. Firstly, the text data in the dataset is processed as through syntactic features: n-grams and distributional semantic derived. The most appropriate n-gram and n-gram range types for each dataset were explored through grid search. The

designer selected SVM and LR frame for building the model. The model will consider the data as a binary yer-or-no decision (hate speech or not hate speech).

3.3. Enhancing the robustness of the model by simulating adversarial attack

As mentioned earlier, some users are currently trying to make changes to some words, making the system unable to understand their text, which is also known as a character-level adversarial attack. TEXTBUGGER is a framework that can simulate this type of change. To improve the robustness of the model, the designers improve the dataset, applying scalable simulation methods in the text [7]. The model is based on the LSTM and combined with CNN and attention mechanisms. The framework for processing the data:

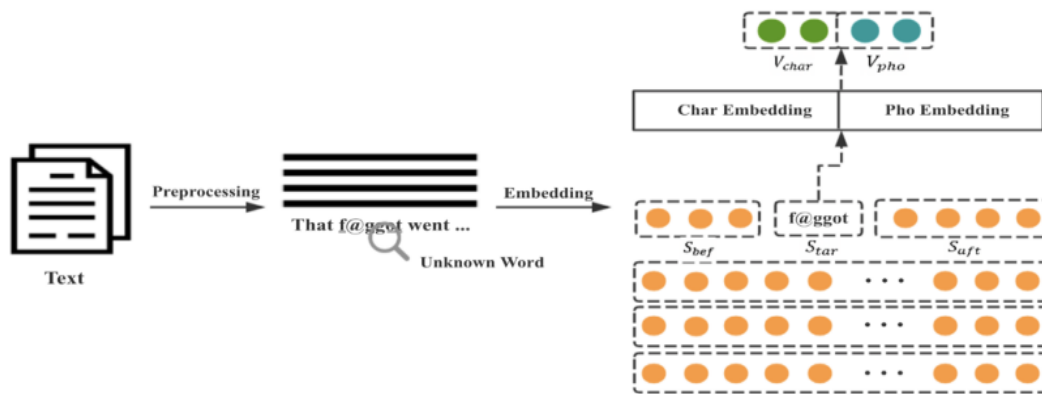


Figure 2. The framework for processing the text for the model

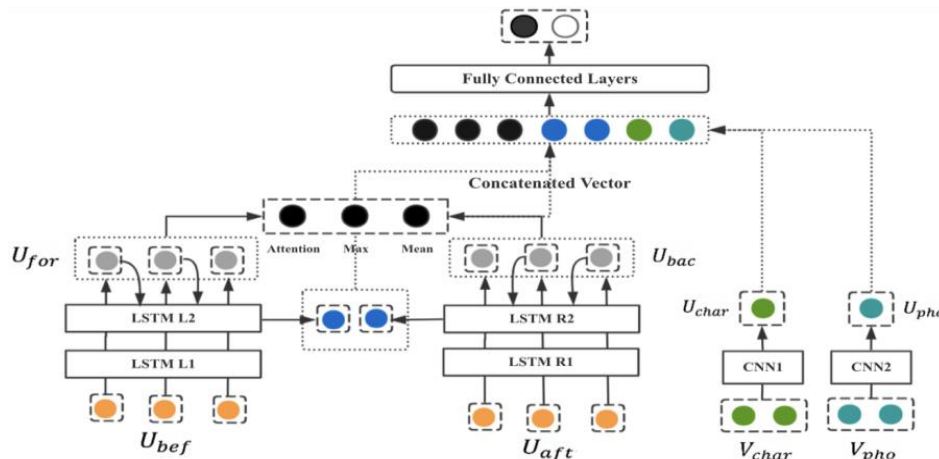


Figure 3. Model for analyzing the text

The text content would be embedding in the progress: Phonetic-level embedding and Word-level embedding. These data would be used for training the model.

4. Challenges

As mentioned above, the current speech environment is more complex, and there is a more complex definition of whether speech is hatred. For example, some groups may have similar contexts and repeat discriminatory words in the complaints they have encountered. Researchers usually use natural language analysis models to analyse context and solve such problems. In order to improve the accuracy of detection models, researchers should always pay attention to the latest natural language analysis models.

So far, there is no unified, continuously updated and widely classified data set available for use. The most frequently used dataset was published in 2016, and the number of classifications is too small. In order to improve the accuracy of hate speech detection, researchers used as many different

datasets as possible. However, the classification of these data sets is different from each other, which leads to the researchers can only classify the data in the data set according to whether it is hate speech or not. Therefore, the trained model only has the dichotomy ability; moreover, there is no unified data set for modelling, which makes it impossible to compare the models with each other. Therefore, in order to further improve the ability and accuracy of the detection model and compare the differences between the research models, a standard, comprehensive and continuously updated dataset is needed for researchers to use uniformly. What's more, many researchers have considered texts in different languages, but there are currently very few non-English hate speech datasets, which are inconvenient for further research.

Different countries and different platforms have different rulers for managing hate speech. Such as, hate speech in the United States is part of civil liberties and only the promotion of violence and threats is illegal; in China, the publication of hate speech is already illegal. Similarly, Hateful humor on Facebook is not illegal, but on bilibili, the largest video site in China, such videos and comments have been completely deleted (Bilibili, 2022). Therefore, if the researchers need to train a hate speech detection model, they should build a model suitable for the current environmental policy, to avoid unnecessary problems.

At present, the expression of hate speech is further upgraded. Some people will use pictures, videos and other forms to publicize hate speech, which requires other types of models to solve such problems.

5. Conclusion

Through sorting out the existing research results, we analysed the advantages of introducing deep learning to detect hate speech in the current language environment, sorted out the definition of hate speech and the types of hate speech, introduced the data sets and evaluation indicators commonly used in the current training model for hate speech detection. Finally, the difficulties and problems that should be paid attention to are put forward.

References

- [1] Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., & López, H. M. H. (2021). Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior*, 58, 101608.
- [2] Li, H. (2017). Deep learning for natural language processing: advantages and challenges. *National Science Review*.
- [3] MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PloS one*, 14(8), e0221152.
- [4] Ayo, F. E., Folorunso, O., Ibharalu, F. T., & Osinuga, I. A. (2020). Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review*, 38, 100311.
- [5] Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4), 1-30.
- [6] Poletto, F., Basile, V., Sanguinetti, M. et al. Resources and benchmark corpora for hate speech detection: a systematic review. *Lang Resources & Evaluation* 55, 477–523 (2021). <https://doi.org/10.1007/s10579-020-09502-8>.
- [7] Guanyi Mou, Pengyi Ye, and Kyumin Lee. 2020. SWE2: SubWord Enriched and Significant Word Emphasized Framework for Hate Speech Detection. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 1145–1154. <https://doi.org/10.1145/3340531.3411990>.
- [8] P. K. Roy, A. K. Tripathy, T. K. Das and X. -Z. Gao, "A Framework for Hate Speech Detection Using Deep Convolutional Neural Network," in *IEEE Access*, vol. 8, pp. 204951-204962, 2020, doi: 10.1109/ACCESS.2020.3037073.
- [9] Unsvg, E. F., & Gambck, B.. (2018). The Effects of User Features on Twitter Hate Speech Detection. *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*.

- [10] Tran, T., Hu, Y., Hu, C., Yen, K., & Park, S.. (2020). HABERTOR: An Efficient and Effective Deep Hatespeech Detector. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).
- [11] G. Mou and K. Lee, "An Effective, Robust and Fairness-aware Hate Speech Detection Framework," 2021 IEEE International Conference on Big Data (Big Data), 2021, pp. 687-697, doi: 10.1109/BigData52589.2021.9672022.
- [12] Facebook, Hate Speech, 2022, available at: <https://transparency.fb.com/zh-cn/policies/community-standards/hate-speech/>.
- [13] Twitter, Hateful conduct policy, 2018, available at: <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>.
- [14] S. Wofson, Facebook labels declaration of independence as 'hate speech', 2018, available at: <https://www.theguardian.com/world/2018/jul/05/facebook-declaration-of-independence-hate-speech>.
- [15] Kinney, Terry A. "Hate Speech and Ethnophaulisms". The International Encyclopedia of Communication. 2008, doi:10.1002/9781405186407.wbiech004. ISBN 9781405186407.