Appliance of Deep Learning on hate speech detection: a systematic review

Dasheng Lin*

Department of Computer Science, Case Western Reserve University, Ohio, United States

*Corresponding author: dxl746@case.edu

Abstract. Hate speech is defined as offensive viewpoints toward individuals or groups divided by ethnicity, religion, or gender. Hate speech is not restricted by law in many countries, and may be seen as the protection of free speech since whether a statement is hate speech is highly related to the individual. Hate speech is widely recognized and has been researched extensively in recent years due to its nebulous definition and quick spread. With deep learning's ongoing growth, some researchers have started to apply deep neural networks to hate speech detection (HSD). Although a lot of progress has been made in research-based work around the task, a comprehensive overview of the progress of the task is lacking. Therefore, an overview of the progress that using deep neural networks to address hate speech problems is necessary. The overview presents what methods have been employed recently to enhance the task's performance, and also analyzes the potential problems of the task.

Keywords: Hate speech; Deep Learning; Offensive languages.

1. Introduction

According to the definition of the Cambridge Dictionary, hate speech is defined as speaking out against someone or a group in public while inciting violence because of their race, religion, sex, or sexual orientation [1]. However, the definition of hate speech and its penalties in law varies among countries. The debate over hate speech, hate speech legislation, and free speech continues to rage [2]. Some national laws describe hate speech as statements, acts, or words that promote stereotypes or violence towards a group or persons with sensitive membership. Hate speech is regarded to be free speech in several nations, including the United States, and is therefore legally protected. As a result of messages being transmitted and received nearly instantly, major social media networks (SMNs) in these nations have now replaced traditional media as the principal means of spreading HS.[3] Because of its complexity and broad presence, hate speech on SMNs becomes a difficult task to settle. Several internet entrepreneurs such as Google, Yahoo, and YouTube mark high priority on this issue to improve their user experience. For example, YouTube did not allow any type of hate speech promoting violence toward groups or individuals divided by ‘age’, ‘disability’, ‘race’, ‘nationality’, ‘religion’, ‘sex/gender’, etc. Twitter's mission is to safeguard its users from violent threats and wishes, hopes, or calls for significant harm on a person or group of people, etc.

In the early stage, the HSD is mainly performed by using machine learning methods: Firstly, the features of the task are constructed manually, and then some machine learning algorithms are used to classify or regress the task. Various classifiers are selected to address the task such as K-Nearest Neighbors (KNN), Decision Trees (DT), and Support Vector Machines (SVM). However, machine learning algorithms have several problems: firstly, the quality of the features is unstable because the constructed task features depend on individual knowledge; secondly, these features are shallow, mostly obtained based on the statistical method which is unable to suitably portray the deep semantic information. Therefore, researchers make use of Deep Learning (DL) to replace Machine Learning (ML) and apply it to the hate speech detection task. By using DL over ML, these advantages are reflected: 1) Shallow ML heavily relies on successful manual extracted features. It serves as the foundation for the various model-building techniques used by each type of learning algorithm. In contrast, DL uses its capabilities for automated feature learning to generate models using high-dimensional raw data as input. That makes DL much more convenient and universal than ML. [4] 2)

Traditional ML process often consists of several independent modules, namely, non-end-to-end learning. For example, a typical Natural Language Processing (NLP) problem could be divided into several independent steps such as word division, lexical annotation, syntactic analysis, etc. The result of every single step would affect the next step and thus the whole training result. Nevertheless, DL adopts end-to-end learning that compared the predicted result with the real result that the model ends up with and produces an error. This error will be used for the adjustment of each layer of the model (such as backpropagation). End-to-end learning would not over until the model convergence or achieves the expected effect. By Difference Theory, end-to-end learning reduces the error propagation compared to traditional machine learning as it optimizes the overall number of steps. DL can automatically learn semantic information about a task through multi-layer nonlinear transformations. Therefore, some researchers have applied Deep Learning to the process of hate speech in recent years.

Despite the fact that Deep Neural Networks have been used to solve such issues and improve the performance of the task, However, reviews about this task are deficient. Many beginners approach the task without being well informed about the latest progress of the task and what latest approaches have been proposed. Besides, a review could be seen as a stage summary that could help in the development of subsequent research-oriented work. Therefore, this paper have reviewed and summarized the development of the task in recent years. This paper first distinguish the definition of the task and then introduce the common datasets and evaluation metrics of the task. In addition, this paper also explain the mainstream deep learning methods used in the task. Finally, this paper introduce some of the latest methods and potential challenges in the tasks.

2. Background

2.1. Task definition

Hate speech, a definition known for its duality, represents the tipping point between free speech and infringement of others, exponentially growing with the development of communication technology. As one's words become much readily known to the public as first-hand information or second-hand information, there are more subjective interpretations of the same sentence as one's interpretation would vary by one's race, gender, sexual orientation, religion, etc... HS is a difficult notion for both people and robots to understand since it has many facets and is complicated, both by humans and machines, as users may adopt different languages, offensive words with similar consonants, or characters substitution in offensive words to bypass the auditor. HS suffers from the problem of determining whether a sentence contains aggressive or dangerous language.

2.2. Datasets

SLEBW [5]: 5759 extracted comments from two sites. Annotated with the following three labels: 'racist', 'non-racist', and 'invalid' by a total of three annotators X, Y, and Z. X and Y are students with similar ages and backgrounds, and Z acts as a tiebreaker when X and Y did not agree on a label. Agreement is measured as Precision (P = 0.49), Recall (R = 0.43), F-measure (F = 0.43).

CJAYY [6]: 2 million comments came from three Yahoo! News and Yahoo! Finance corpora. Annotated as "clean" or "abusive" by Yahoo internal software. These annotated comments are used to train a classifier that is then applied to 1.1 million annotated comments from a different dataset. Finally, a third, smaller dataset is utilized for evaluation, and the comments are classified as either "abusive" or "clean" by three qualified auditors. Fleiss' κ coefficient is 0.843, and the agreement rate is 0.922.

TDMI [7]: 25,000 randomly chosen tweets from 33,458 Twitter users sample including 85.4 million tweets, categorized into one of three classifications by CrowdFlower staff: hate speech, offensive but not hate speech, or neither offensive nor hate speech. A team of three or more employees coded each tweet. Agreement is measured as Precision (P = 0.91), Recall (R = 0.9), F-measure (F = 0.9).

FAFMM [8]: 99 articles with a total of 17,567 comments on Facebook were crawled from the chosen sites. Any of the three levels of hatred—No hatred, Weak hatred, and Strong hatred—is annotated. Five bachelor students would annotate each comment a minimum of once and a maximum of five times. Religion, physical and/or mental disability, social class, politics, racism, sex and gender issues, and others are some of the areas that hate speech falls under. Measured by agreements is the degree to which various annotators agree on a job, resulting in Fleiss' κ (κ = 0.19) when discriminating on three hate classes and Fleiss' κ (κ =0.26) over Strong Hate and Weak Hate, and Precision (P = 0.833), Recall (R = 0.872), F-measure (F = 0.851).

WH [9]: 136,052 retrieved tweets and 16,914 annotated tweets over two months. According to the author, A tweet is offensive if one fits any of the below 11 descriptions: utilizes a racial or gendered epithet; minority assault; aims to intimidate a minority; condemns a minority (without a well-founded argument); supports horrific crime or hate speech but does not really engage in any of those things; etc. The author gave inter-annotator agreement Fleiss' κ (κ = 0.84). Argument is measured as Precision (P = 0.72), Recall (R = 0.77), F-measure (F = 0.73).

2.3. Metrics

In this section, this paper would introduce metrics leveraged in the HSD task. One could refer to these formulas to select an applied database.

$$P = \frac{|\text{relevant documents} \cap \text{retrieved documents}|}{|\text{retrieved documents}|} \tag{1}$$

Eq.1 has displayed the definition of Precision(P) in the field of information retrieval, where relevant documents could be seen as the number of hate speech and retrieved documents could be seen as the total number of comments in a dataset under this circumstance.

$$R = \frac{|\text{relevant documents} \cap \text{retrieved documents}|}{|\text{relevant documents}|} \tag{2}$$

Eq.2 has displayed the definition of Recall(R). the fraction of the relevant document that was successfully retrieved. However, it is somehow meaningless to achieve a high rate of recall as returning all documents could easily reach. Henceforth, both P and R should use together, as calculating P means measuring the number of non-relevant documents.

$$F = 2 \times \frac{P \times R}{P + R} \tag{3}$$

Eq.3 has displayed the definition of F-measure(F) that combines P and R. F is also known as the balanced F-score, F1 measure, or the standard F-measure, where the harmonic mean of P and R is represented by F as the weight of P and R are evenly equal.

$$\kappa = \frac{\overline{P} - \overline{P_e}}{1 - \overline{P_e}} \tag{4}$$

Formula.4 gives the definition of Fleiss' κ, a metric for inter-rater reliability in statistics, a generalization of Scott's π statistic.[10] If all things are distributed equally and randomly to each rater, Fleiss' κ is applicable for any number of raters categorizing any number of objects. A more intuitive understanding is that Fleiss' κ could measure the consistency of ratings when these raters are asked to give numerical indicators to items, for example, determining a sentence is 'abusive'(-1) or 'clean'(0). κ = 1 if all raters have the same agreement or give the same numerical indicator, and κ = 0 if there is no agreement among these raters.

3. Recent Researches

In this section, we give a quick summary of a few recent studies that have used deep learning rather than machine learning to study hate speech, presenting the advantages of deep learning over traditional machine learning.

3.1. Hate Speech Detection with Comment Embeddings

The author discusses the issue of recognizing hate speech in user comments online that are expressed as hostile sentiments toward particular groups that are split by race, religion, or gender.[11]. Based on the internet ‘s non-restrictive nature and protection of free speech, hate speech is widely published on social media platforms. These users promote simple and effective techniques to avoid detection, such as removing or replacing keywords and phrases, while their content remains offensive to human readers. Due to this, the problem is highly dimensional and highly sparse, leading to the overfitting of the model. To address these issues, the author proposes a two-step method. First, the continuous BOW neural language model is utilized in conjunction with paragraph2vec, a recently published neural language model, to jointly represent comments and sentences. The author thus trains a binary classifier by using embeddings to eliminate hate speech. One may evaluate the author’s method using the largest data set available for the relevant study, which consists of 895,456 clean comments and 56,280 comments that detecting hate speech that was obtained from the Yahoo Finance website. Using paragraph2vec, the author obtains the dominant Area under the Curve (AUC), while paragraph2vec requires less memory and training time, making this method one of the most competitive methods.

3.2. Hate-Speech and Offensive Language Detection in Roman Urdu

Hate Speech is a huge and specific problem in Pakistan [12]. Pakistan recognizes Urdu as its national language even though English is the country's official language. Compared to formal languages, RU is much harder to represent since people frequently transition between two languages in the same discussion. Although Roman Urdu is a widely spoken language, the author discusses how limited the study on hate speech is because annotated datasets, language models, and language resources are lacking. With binary coarse-grained and multi-class fine-grained labels, they produce the annotated data set RUHSOLD from tweets in RU, determine if it is feasible to teach RU language the learning of five current embedding models with numerous tests, introduce the CNN-gram, an unique deep learning architecture, and more than 4.7 million tweets were used to teach domain-specific embeddings, RomUrEm, to exhibit. They also present a list of 621 hateful words in RU. With seven baselines on the RUHSOLD data set, CNN-gram demonstrated higher resilience in both coarse- and fine-grained classification tasks. RomUrEm also reported high performance compared to five existing embedding models while BERT acts similarly to RomUrEm under some circumstances.

3.3. Hierarchical CVAE for Fine-Grained Hate Speech Classification

Automatic models for detecting various hate speech have been developed in response to the large growth of online hate [13]. About four out of ten Americans say they have been the victim of internet harassment, and 63 percent say it is a big problem, as reported by PewResearch Center. The author does point out that both detection models share the issue of concentrating on categorizations that are coarse-grained and have a few categories, which frames this challenge as a classification issue with less than seven types. Pioneering, the author examines fine-grained hate speech categorization that links specific tweets to hate groups. In order to create a model for more precise classification, the author suggests grouping the tweets from 40 hate organizations into 13 separate categories. By observing the robustness of CVAE with decreasing data sets, in order to complete this objective, the author expands and creates a Hierarchical CVAE model, leveraging extra data on hate groups as a training tool. According to experimental evidence, the author demonstrates that using data from the hate categories for training can greatly enhance classification accuracy, as this model raises baseline

Micro-F1 scores by up to 10%, compared to models such as CVAE, Bi-LSTM when 90% of the entire data set is used for training.

3.4. Hate Speech Detection based on Sentiment Knowledge Sharing

Automatic detection and active avoidance mechanism for hate speech. However, the existing method for hate speech detection suffers from inherently biased training because of stereotyped words [14]. For example, the sentence is judged as offensive if the second person pronoun and pejorative are present at the same time. The author claims that artificial features such as a variety of linguistic rules and dictionaries of derogatory words make the performance of neural networks unsatisfactory, address the problem, and thus offer a framework for detecting hate speech that is based on sentiment knowledge sharing. Instead of detecting words with strong negative emotions, the author uses multiple feature extraction units and use a gated attention mechanism to fuse features to achieve these achievements: 1) the author not only integrates the negative word of target sentences but also uses multi-task learning to share external sentiment knowledge and to lead the learning of model. 2) The author suggests a brand-new structure with several feature extraction modules. Each extraction unit extracts features using a feedforward neural network and a multi-head attention mechanism, then employs gated attention to fuse features. Experimentally, the author's framework achieves advanced performance compared to strong baselines. Further detailed examples also check the model's performance in detecting hate speech.

4. Challenges

As we have mentioned, hate speech is a complex phenomenon. Although researchers contributed significantly to the study of hate speech specifically in recent years, there are still many difficulties in this area:

Researchers may show great divergence in the agreement of whether one sentence is hated speech or not when creating a baseline database. Even the same sentence might have different meanings under different circumstances. This is sometimes related to the context in which the phrase is used.

Due to the quick development of society, culture, and language, especially internet language, evolves also quickly. That causes the database to fail faster and faster, A model trained using a database that is only a few years old is likely to be unsuitable for the current network environment. Error caused by this might pass to research that is based on this, and data from a few years ago in this database may cause an over-fitting problem.

Hate speech and other kinds of offensive speech, such as abusive speech, can easily be conflated, as these speeches can be easily mixed. Especially in end-to-end learning in Deep Learning, if the distinction between hate speech and abusive speech is not obvious, the model is somehow prone to errors.

5. Conclusion

This paper examine the benefits and drawbacks of utilizing deep learning to distinguish hate speech from the huge amount of data such as comments, tweets, blogs, etc. in this article. This paper thus summarize some researches that adopt Deep Learning to distinguish hate speech, list the Deep Learning methodology they used, and the performance they obtained compared to the baseline. Based on the current research, future research could start with local dialects / non-national languages that are heavily used on the network but not well known, or further research on languages with high semantic toughness (e.g., simple homophone replacement, deletion, abbreviation, or rewriting of offensive words that leave the aggressive content unchanged).

References

[1] Dictionary.cambridge.org.

[2] Herz, M. E., & Molnár Péter. (2012). The content and context of hate speech: Rethinking regulation and responses. Cambridge University Press.

[3] Mullah, N. S., & Zainon, W. M. (2021). Advances in machine learning algorithms for hate speech detection in Social Media: A Review. IEEE Access, 9, 88364–88376. https://doi.org/10.1109/access.2021.3089515

[4] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature. https://doi.org/10:1038

[5] Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A dictionary-based ap- proach to racism detection in dutch social media. arXiv Preprint arXiv:1608.08738 (2016)

[6] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 145–153.

[7] Thomas D., Dana W., Michael M., and Ingmar W. 2017. Automated hate speech detection and the problem of offensive language. arXiv Preprint arXiv:1703.04009 (2017).

[8] Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. In NLP4CMC III: 3rd workshop on natural language processing for computer-mediated communication.

[9] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'16). 88–93).

[10] Gwet, K. L. (2014) Handbook of Inter-Rater Reliability (4th Edition), Chapter 6. (Gaithersburg : Advanced Analytics, LLC) ISBN978-0970806284.

[11] Labs, N. D. Y., Djuric, N., Labs, Y., Yahoo!, J. Z., Zhou, J., Yahoo!, Yahoo!, R. M., Morris, R., Labs, M. G. Y., Grbovic, M., Labs, V. R. Y., Radosavljevic, V., Labs, N. B. Y., Bhamidipati, N., Council, N. R., Rome, S. U. of, & Metrics, O. M. V. A. (2015, May 1). Hate speech detection with comment embeddings: Proceedings of the 24th International Conference on World Wide Web. ACM Other conferences. Retrieved August 24, 2022, from https://dl.acm.org/doi/abs/10.1145/2740908.2742760?casa_token=Kj6Qx1Yr_JIAAAAA%3AhprSMaotYu6NFYEeffluiI4Sk3-COURs8t6Bh8KzUHDP--03CIGFiN97zAaXh5JJrQej5au4p6X5zQ

[12] Rizwan, H., Shakeel, M. H., & Karim, A. (n.d.). Hate-speech and offensive language detection in Roman Urdu. ACL Anthology. Retrieved August 24, 2022, from https://aclanthology.org/2020.emnlp-main.197/

[13] Qian, J., ElSherief, M., Belding, E., & Wang, W. Y. (2018, August 31). Hierarchical CVAE for fine-grained hate speech classification. arXiv.org. Retrieved August 24, 2022, from https://arxiv.org/abs/1809.00088

[14] Zhou, X., Yong, Y., Xiaochao Fan, Ren, G., Song, Y., Diao, Y., Yang, L., & Lin, H. (n.d.). Hate speech detection based on sentiment knowledge sharing. ACL Anthology. Retrieved August 24, 2022, from https://aclanthology.org/2021.acl-long.556/