engineering proceedings

MDPI

Proceeding Paper

Hate Speech Detection: Performance Based upon a Novel Feature Detection †

Saugata Bose

School of Computing and Information Technology, Faculty of Science and Engineering and Information Sciences, University of Wollongong, Wollongong, NSW 2500, Australia; saugata28@gmail.com

† Presented at the 3rd International Electronic Conference on Applied Sciences, 1–15 December 2022; Available online: https://asec2022.sciforum.net/.

Abstract: Hate speech is abusive or stereotyping speech against a group of people, based on characteristics such as race, religion, sexual orientation, and gender. Internet and social media have made it possible to spread hatred easily, fast, and anonymously. The large scale of data produced through social media platforms requires the development of effective automatic methods to detect such content. Hate speech detection in short text on social media has become an active research topic in recent years, as it differs from the traditional information retrieval for documents. My research is to develop a method to effectively detect hate speech based on deep learning techniques. I have proposed a novel feature based on the lexicon of short text. Experiments have shown that proposed deep-neural-network-based models improve the performance when a novel feature combines with CNN and SVM.

Keywords: hate speech; CNN; SVM; feature detection

check for updates

Citation: Bose, S. Hate Speech Detection: Performance Based upon a Novel Feature Detection. *Eng. Proc.* **2023**, *31*, 87. https://doi.org/10.3390/ASEC2022-13788

Academic Editor: Nunzio Cennamo

Published: 2 December 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Internet connectivity allows people to express their opinions. They are doing this either by writing blogs, which have no limitation on length, or by posting status updates and comments on social media, which have very specific, though varying length constraints. Like Facebook, it allows 63,206 characters in status updates, whereas Twitter allows only 280 characters in status updates [1]. Nowadays, social media becomes a crucial part of human life. Our lives have started getting controlled somewhat by those published statuses or comments from the "owner" of those statuses or comments hiding behind in both positive and negative ways. The International Covenant on Civil and Political Rights (ICCPR) states that "any advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility, or violence shall be prohibited by law". My research is to analyze those posts or comments published in social media to analyze their semantics, but I restrict this study to detecting "hate speech" from short text on social media, specifically Twitter. I propose an approach based on feature extraction to extract relevant features for short text. The performance of hate speech detection is highly dependent on the features used to characterize the hate speech. Literature review found that several application-dependent features built by experts, such as parts-of-speech, etc., have been proposed. Most proposed features in the literature do not perform very well. I argue that the poor performance of those features is rooted in their weak correlation with the semantics of the speech. In this study, I have proposed a feature based on the hate speech lexicon.

2. Proposed Method

Adding manually picked features to the deep trained features will improve the accuracy score of the classification after feeding to a nonlinear SVM classifier (see Figure 1). In this study, I experiment with a feature which is lexicon-based. This feature will tell us how

much hate or non-hate a tweet has, by looking at the presence of the 'hated' and 'non-hated' words. These words have strong correlation with the tweet label. If X is a dataset having n number of tweet documents $\{X_1, X_2, \ldots, X_n\}$, with categorical class labels for hate in $d1$ number of hated tweets, and for non-hate in $d2$ number of non-hated tweets, then T would contain m number of non-hate words $\{T_1, T_2, \ldots, T_m\}$, which have strong correlation with the label non-hate, and U would contain p number of hated words $\{U_1, U_2, \ldots, U_p\}$, which are related to the hated class.

The frequency of the non-hate words appearing in the non-hate tweets would give us a notion of how much weight a particular non-hate word carries in the specific tweet. If T_i, where $i = 1$ to m appears C_i, where $i = 1$ to an integer number of times in the $d1$ number of documents, then F_i would illustrate the weight of each non-hate words in the non-hate tweet document set.

$$F_i = \frac{C_i}{d1} \tag{1}$$

$$\text{and } G_i = \frac{D_i}{d2} \tag{2}$$

A similar equation can be formed for calculating the weight value of the hated words in the $d2$ number of hated document sets. The cumulative of F_i in a specific non-hate tweet document refers to the weight of the non-hate words in the tweet. For any presence of non-hate words, $T_i = \{T_1, T_2, \ldots, T_m\}$ in a $Z_i = \{Z_1, Z_2, \ldots, Zd_1\}$ non-hate tweet, the weight value of the non-hate tweet would be the cumulative sum of the presence of the F_i.

$$\forall T_i \in T,\ weight\ of\ Positive\ Tweet_i = \sum_{i=1}^{d1} F \tag{3}$$

In this way, the weight of each hate tweet can be calculated, as well by the following equation,

$$\forall U_i \in U,\ weight\ of\ Negative\ Tweet_i = \sum_{i=1}^{d2} G \tag{4}$$

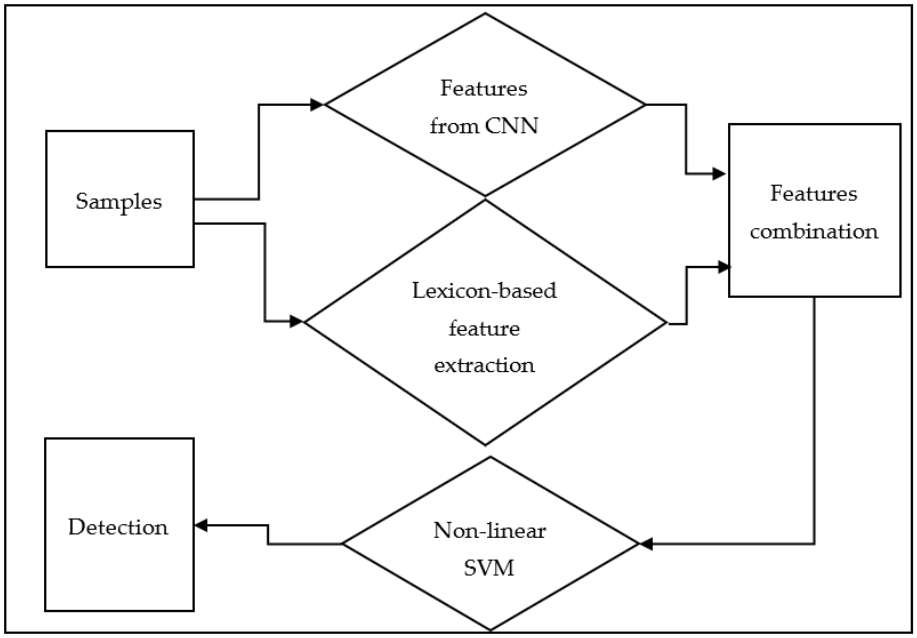


Figure 1. Proposed Model. Combine features (Features extracted from CNN model and manual features) are fed to a nonlinear SVM classifier to detect "hate speech".

This is a unique feature which has never been experimented on by researchers. This weight value tells us how much hate or non-hate sentiment a tweet carries.

I integrated this feature with the outputs extracted from a CNN model and fed them to an SVM classifier. Then, the SVM classifier had enough features to get trained and create the margin.

3. Results

In this section, the performance of the proposed model has been presented. Accuracy score, recall score, precision score, and F1 score have been used as performance metrics. We concentrate these performances on the "hate class" label only. The model has been trained with unbalanced and balanced data. The Table 1 projects a high average F1 score (i.e., 0.97) on an unbalanced dataset compared to the previous studies reported in Table 1. Moreover, we present the macro-F1 score on the balanced dataset, which shows a significant improvement compared with [2]. In summary, we could conclude that the proposed method has a significant possibility for "hate speech" detection.

Table 1. Performance of the proposed classifier on test data. We use accuracy, micro-F1 and macro-F1 as performance metrics. The table demonstrates the model's performance on balanced and unbalanced datasets [2]. Moreover, the performance on [2] was compared with previous researchers' findings. The best results are highlighted in bold and underlined.

Dataset Type	Accuracy	Recall	Recall-Hate Class	Precision	Precision-Hate Class	Micro-F1	Macro-F1
Unbalanced	0.936755	0.981986	0.199301	0.952371	0.404255	**0.966952**	0.266979
Balanced	0.702797	0.730769	0.674825	0.692053	0.714815	0.710884	**0.694245**
Performance Reported on Davidson dataset [1] by Previous Studies							
Dataset				**Scores**			
[2]		0.61		0.44		0.90	
[3]						0.94	0.30
[4]		0.89		0.89		0.89	
[5]						0.924	

4. Conclusions

The paper has proposed a novel framework for hate speech detection. Through comprehensive experiments, we found that the proposed method was able to detect unique lexicon-based features from short-text "hate speeches". Integrating these feature sets with a deep learning framework creates a novel framework that enhances the model's performance.

Supplementary Materials: The material is available at https://www.mdpi.com/article/10.3390/ASEC2022-13788/s1, Poster: Hate Speech Detection: Performance Based upon a Novel Feature Detection.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is available at https://github.com/t-davidson/hate-speech-and-offensive-language (10 November 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Know Your Limit: The Ideal Length of Every Social Media Post. Available online: https://sproutsocial.com/insights/social-media-character-counter/#facebook (accessed on 10 November 2022).
2. Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. Automated Hate Speech Detection and the Problem of Offensive Language. In Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM-17), Montreal, Canada, 15–18 May 2017; pp. 512–515.
3. Zhang, Z.; Robinson, D.; Tepper, J. Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In *Lecture Notes in Computer Science, Proceedings of the 15th Extended Semantic Web Conference, ESWC'18, Heraklion, Greece, 3–7 June 2018*; Springer: Cham, Switzerland, 2018; Volume 10843, pp. 745–760.
4. Founta, A.M.; Chatzakou, D.; Kourtellis, N.; Blackburn, J.; Vakali, A.; Leontiadis, I. A Unified Deep Learning Architecture for Abuse Detection. Proceedings of the WebSci '19: 10th ACM Conference on Web Science, Boston, MA, USA, 30 June 2019–3 July; pp. 105–114.
5. Kshirsagar, R.; Cukuvac, T.; McKeown, K.; McGregor, S. Predictive Embeddings for Hate Speech Detection on Twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 26–32.