ASEC 2022

*Abstract*

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

**Saugata Bose**

[1] University of Wollongong, Australia; saugata28@gmail.com

† Presented at the 3rd International Electronic Conference on Applied Sciences; 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 becomes an active research topic in recent years as it differs from 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 lexicon for short text. Experiments have shown that proposed deep neural network based model improves performance when novel feature combines with CNN and SVM.

**Keywords:** keyword 1; keyword 2; keyword 3

**Citation:** Lastname, F.; Lastname, F.; Lastname, F. Hate Speech Detection: Performance Based upon a Novel Feature Detection. *Eng. Proc.* **2022**, *4*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s):

Published: 1 December 2022

I assume that if manually picked features are added to the features extracted from the trained network, it may improve the accuracy score of the classification. In this study, I experiment with a feature which is lexicon based. This feature will tell us how much hated or non-hate a tweet is by looking at the presence of the 'hated' and 'non hated' words. These words have strong co relation 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-hate tweets. Then $T$ would contain m number of non-hate words $\{T_1, T_2, \ldots, T_m\}$ which have strong co relation 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 that how much weight a particular non hated words 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 hated words in the non-hate tweet document set.

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

$$and\ G_i = \frac{D_i}{d2}$$

Similar equation can be formed for calculating the weight value of the hated words in the *d*2 number of hated document sets. The cumulative of $F_i$ in a specific non hated tweet document refers 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 hated tweet, the weight value of the non-hate tweet would be the cumulative sum of the presence of the $F_i$

$$\forall T_i \in T, weightOfPositiveTweet_i = \sum_{i=1}^{d1} F$$

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

$$\forall U_i \in U, weightOfNegativeTweet_i = \sum_{i=1}^{d2} G$$

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

I integrate this feature with the outputs extracted from a CNN model and fed these to an SVM classifier. Then the SVM classifier has enough feature to get trained to create the margin

**Author Contributions:**

**Funding:**

**Institutional Review Board Statement:**

**Informed Consent Statement:**

**Data Availability Statement:**

**Conflicts of Interest:**

## References

1. Author 1, A.B.; Author 2, C.D. Title of the article. *Abbreviated Journal Name* **Year**, *Volume*, page range.