ISSN 2007-9737

Automatic Hate Speech Detection Using Deep Neural Networks and Word Embedding

Olumide Ebenezer Ojo[1], Thang-Hoang Ta[1,2], Alexander Gelbukh[1], Hiram Calvo[1], Grigori Sidorov[1], Olaronke Oluwayemisi Adebanji[1]

[1] Instituto Politécnico Nacional, Centro de Investigacion en Computación, Mexico

[2] Dalat University, Lam Dong, Vietnam

olumideoea@gmail.com, thangth@dlu.edu.vn, gelbukh@gelbukh.com, hcalvo@cic.ipn.mx, sidorov@cic.ipn.mx, olaronke.oluwayemisi@gmail.com

Abstract. Hatred spreading through the use of language on social media platforms and in online groups is becoming a well-known phenomenon. By comparing two text representations: bag of words (BoW) and pre-trained word embedding using GloVe, we used a binary classification approach to automatically process user contents to detect hate speech. The Naive Bayes Algorithm (NBA), Logistic Regression Model (LRM), Support Vector Machines (SVM), Random Forest Classifier (RFC) and the one-dimensional Convolutional Neural Networks (1D-CNN) are the models proposed. With a weighted macro-F1 score of 0.66 and a 0.90 accuracy, the performance of the 1D-CNN and GloVe embeddings was best among all the models.

Keywords. Hate speech, gloVe, 1D-CNN.

1 Introduction

The development of social media and other networking sites enables people to discuss and express themselves. This can be highly beneficial and can also pose tremendous danger to the peace of the society. Analysing sentiments in social media text is an important field of natural language processing and an integral part of many applications. Text from social media sites can contain many risks such as hate speech, fake news, violence, intimidation, racism and can also be life-threatening at times [23, 2, 11, 3, 1, 4, 17].

Individuals often share various opinions that can lead to unhealthy and unequal debate. Fostering a healthy dialogue is difficult for various social media sites, with these platforms being forced either to restrict or shut down users' comments. This study focuses on developing a model for detecting hate speech in English text on an internet forum site.

Humans discriminate against others based on their affiliations, classifying them as belonging or not belonging to a shared identity. The freedom to post online has prompted users to communicate differently, which can often lead to controversial outtakes on other individuals or on specific issues. Occupying different positions in politics and business is making communication play different roles in various events. Communication sites have made several efforts to moderate, but have restricted capacity. Needless to say, these moderators must put their time, resources and effort into managing their platforms against any form of negativity. The goal of this research is to detect hate speech in expressions in an efficient

and accurate manner, as well as to assist in the interpretation of text views.

Different machine and deep learning models was used to classify sentences and to access opinions in Ojo et.al [18]. This offers the edge to examine the perspectives of people on significant economic activities by studying their characters. It is really interesting to analyze these opinions from people about events and various issues as a way of knowing what they are thinking of, planning to do or engaged with at a particular time. In times like this, when decisions and responses are generated and modified in seconds, detecting hate speech in text is extremely necessary. The classification of text on social media platforms can be done using text classification tools [22, 12, 10, 15, 19].

In this paper, we used a deep learning approach to recognize different types of hatred in text and we focused on dataset from posts on a white supremacist forum, Stormfront [5]. We tried out different classification methods known as Naive Bayes Algorithm (NBA), Logistic Regression Model (LRM), Support Vector Machines (SVM), Random Forest Classifier (RFC) and the one-dimensional Convolutional Neural Networks (1D-CNN). The Bag-of-Words (BoW), term frequency–inverse document frequency (TF-IDF), and Global Vectors (GloVe) word embeddings were employed as feature representations.

The sequence classification approach is based on a neural network architecture that, through the representation of words and characters, benefits from the combination of word embeddings and 1D-CNN [3]. Machine learning approaches was also used to learn from the data and to perform the classification task. We used datasets tagged with Hate and No-Hate labels from [5] that was categorized and annotated at the sentence level. We were able to determine which classifier is best to detect hate speech based on the accuracy rate from the models.

2 Literature Review

2.1 General Concept

Deep learning analysis involves the rigorous study of data and requires systematic data analysis. During the analysis, deep connections are established between already existing concepts and new concepts are being developed, allowing long-term retention of ideas so that they can be used in new contexts to solve problems [7]. The architecture of a one dimensional Convolutional Neural Network Models (1D-CNN) and parameters represents a deep learning neural network.

Deep learning approaches like 1D-CNN, have recently been shown to achieve state-of-the-art performance on difficult classification problems [3]. Kernel slides along one dimension in 1D-CNN, which is mostly employed on text and 1D signals. 1D-CNN uses its internal state to process sequential data in order to memorize feature representations, perform classification and prediction tasks, and thus has no conceptual understanding of the data. It combines input vector with state vector to generate a new state vector with a learned function. It allows the measurement of fixed-size vector representations for arbitrary word sequences. We will implement a Convolutional Neural Network type algorithm called 1D-CNN for processing sequential data in this task.

2.2 State of the Arts

Several research works have been conducted using different methods to study and tackle the issue of hate, or toxicity detection in text [20, 9, 14, 1, 2, 5, 3, 4]. Various machine learning approaches, mostly defined by the type of network and training methods, have been used to classify text of this kind. According to [3], hate statements, otherwise known as violent threats, can be likened to a violent crime which affects the individuals or groups targeted. The researchers categorized the threatening comments into those that target an individual or group, and identified the threats of violence. They used a binary classification approach in their work to predict violence threats.

Convolutional Neural Networks (CNN) have also been applied to the text classification task for both distributed and discrete embedding of words [9, 22]. While representations derived from convolutional networks offer some sensitivity to word order, their order sensitivity is restricted to

mostly local patterns, and disregards the order of patterns that are far apart in the sequence. Although some word order sensitivity is given by representations derived from convolutional networks, there is limitation to their order sensitivity which ignores the order of patterns that are far apart in the sequence.

The use of 1D-CNN in [3] yielded better performance in the classification of text. In [18], multiple classifiers such as decision tree classifier, random forest classifier, vector supporting machines, logistic regression model, and others including a deep neural network, were used with n-grams approach to classify the text polarity.

In relation to this work, hate is another term being researched in the scientific community which results to bullying, unrest, embarrassment, and can even cause racism through the use of social media platforms.

Another research carried out on a dataset in [6] have also used the idea of transfer learning during their training process with respect to classification of text from various online communication channels. These networks maintain a state that can retain and reflect data from an indefinitely long text. The network stores several stable vectors in what is known to be memories, which the network remembers when similar vectors are presented to the network memory.

Word embeddings in [8] was used as a machine-learning method to depict each English word as a vector, with these vectors capturing semantic relationships between the associated words. The study looked at how the architecture of word embeddings varies over time and correlates with empirical demographic changes in terms of gender and ethnic stereotypes.

The use of hate speech [5], exist in many similar terms, which includes violent threat [3], offensive behavior [4], language of aggression or abuse [1, 2], and toxicity [14, 9, 20]. We are interested in distinguishing between text that is Hate Speech or not, and the method we proposed is our driving factor in carrying out this study. We will explain our approach in the following sections and provide information about the obtained results.

3 Experimental Analysis

3.1 Data

Gilbert et al. [5] generated a publicly available dataset annotated at the sentence level on Internet forum posts in English. The data were read and individually annotated into two different classes which are Hate and No-Hate. The texts were pre-processed and the embedded terms in text sequence were used as input into the models and the context summarized with a vector representation. We experimented with a number of representations, including BoW and word embeddings with GloVe, and concentrated on machine learning models and a deep-learning classifier. Each of the words in the sentence is then mapped to a (pre-trained) word embedding.

3.2 Methods of Analysis

Machine learning algorithms that were applied to the dataset include NBA, SVM, RFC and LRM. A one dimensional convolutional neural network, 1D-CNN, was also used to learn from the data. We used BoW for the machine learning algorithms and GloVe word embedding techniques for the 1D-CNN. Before performing hate detection classification on the binary label social media text, unnecessary features were removed to increase the efficiency and performance of the classification algorithms.

The data, which had already been labeled, was utilized to determine whether or not a specific text contained hate or no-hate assertions. 70% of the whole dataset was used for training, while the remaining 30% was used for testing. Word vector representations are extremely useful for capturing semantic information. Word2Vec [16] and GloVe [21] are the two most recently described methods for creating word embedding models.

GloVe is more efficient than Word2Vec, according to Pennington et al. [21]. GloVe means Global Vectors, with global referring to global corpus statistics and vectors referring to word representations. To obtain the inputs to the deep learning network, we used the GloVe pre-trained model.

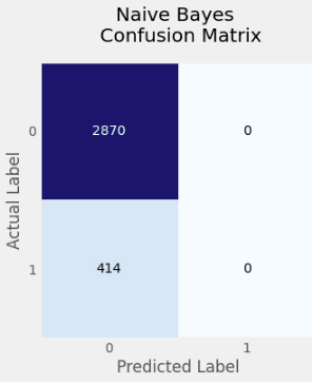

Fig. 1. Confusion Matrix of the NBA Predictions on the Test Set

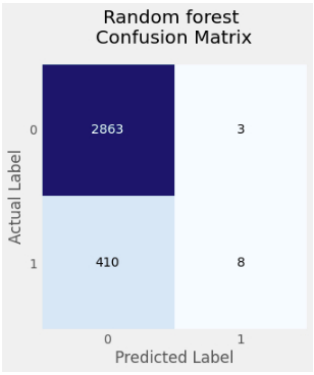

Fig. 2. Confusion Matrix of the RFC Predictions on the Test Set

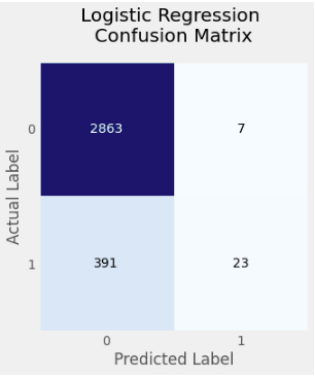

Fig. 3. Confusion Matrix of the LRM Predictions on the Test Set

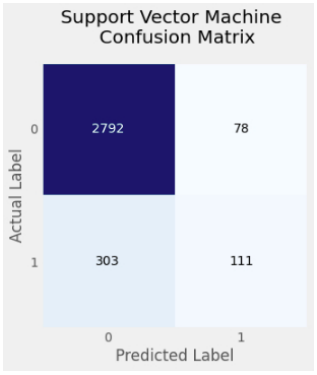

Fig. 4. Confusion Matrix of the SVM Predictions on the Test Set

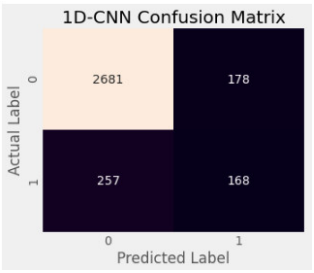

Fig. 5. Confusion Matrix of the 1D-CNN Predictions on the Test Set

A massive corpus of 2B tweets was used to train the GloVe pre-trained model. The machine learning algorithms used the BoW features. We performed optimization using the Adam algorithm [13]. The 1D-CNN learns to encode input sequence properties which are useful for the task of detecting hate speech in the sentence. CNN-based text classifications can learn features from words or phrases in different positions in the text.

4 Results

We classify the data with the models developed. After pre-processing the data, we used the BoW and TF-IDF approaches to convert text sentences into numeric vectors for the machine learning models. Four well-established machine learning algorithms were implemented on the datasets namely SVM, NBA, LRM and RFC. A deep

representation of the words and their relative meanings was done with Glove word embeddings and used by the 1D-CNN deep learning model.

Subsequently, the text was classified and the respective macro averaged F1 scores and accuracy results of all the models are shown in table 1 below. The proposed models for detecting and classifying hate speech in text were evaluated to identify the best algorithm.

Table 1. Accuracy and F1 values for all classification methods

Model	Features	F1	Accuracy
NBA	BoW	0.47	0.81
RFC	BoW	0.49	0.87
LRM	BoW	0.54	0.88
SVM	BoW	0.65	0.88
1D-CNN	GloVe	**0.66**	**0.90**

5 Conclusions

The difficulty of automatically recognizing hate speech in social media posts is addressed in this study. This research presents a hate speech dataset that was manually labeled and collected from a white supremacist online community. We discovered that the analysis generated significant hate preconceptions, as well as ranging levels of ethnic and religious-based stereotypes. Our findings have shown that the selection of word embeddings, the selected parameters and the optimizer have a high impact on the output achieved.

Hate speech in the social media space, which can have negative impacts on the society were detected easily and the high accuracy rate of the model will bring many benefits while reducing the damage. By assessing and comparing the performance of the various hate detection models, we found that word embeddings with 1D-CNN is an important tool for hate speech detection.

1D-CNN, a deep learning model, achieved the highest weighted macro-F1 score of 0.66 with a 0.90 accuracy. The results of the confusion matrix graphs in figures 1 to 5 demonstrated that GloVe embedding features were unable to correctly classify the test dataset. This could be due to the fewer training sentences used byy the GloVe word embedding algorithm. Furthermore, a closer examination of the figures reveal the order in which the models performed best in the dataset.

Acknowledgment

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of the CONACYT, Mexico, and by the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico, under Grants 20211884, 20220859, and 20220553, EDI; and COFAA-IPN. We are grateful to Ona de Gibert Bonet and her colleagues for the dataset.

References

1. **Agrawal, S., Awekar, A. (2018).** Deep learning for detecting cyberbullying across multiple social media platforms. Proceedings of the European Conference in Information Retrieval (ECIR), Grenoble, France, pp. 141–153.
2. **Aroyehun, S. T., Gelbukh, A. (2018).** Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-1), Santa Fe, USA.
3. **Ashraf, N., Mustafa, R., Sidorov, G., Gelbukh, A. (2020).** Individual vs. group violent threats prediction in online discussions using deep learning. Companion Proceedings of the Web Conference 2020, April 20–24, 2020, Taipei, Taiwan, pp. 629–633.
4. **Chen, Y., Zho, Y., Zhu, S., Xu, H. (2012).** Detecting offensive language in social media to protect adolescent online safety. PASSAT 2012, International Conference on Social Computing (SocialCom), IEEE, Amsterdam, Netherlands, pp. 71–80.
5. **de Gibert, O., Perez, N., García-Pablos, A., Cuadros, M. (2018).** Hate speech dataset from a white supremacy forum. Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Association for Computational Linguistics, Brussels, Belgium, pp. 11–20.

6. **Do, C. B., Ng, A. Y. (2005).** Transfer learning for text classification. Advances in Neural Information Processing Systems 18, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada.
7. **Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Rodr´ıguez, J. (2017).** A review on deep learning techniques applied to semantic segmentation. IEEE Transactions on Pattern Analysis & Machine Intelligence, Vol. 39, No. 04.
8. **Garg, N., Schiebinger, L., Jurafsky, D., Zou, J. (2017).** Word embeddings quantify 100 years of gender and ethnic stereotypes. CoRR, Vol. abs/1711.08412.
9. **Georgakopoulos, S. V., Tasoulis, S. K., Vrahatis, A. G., Plagianakos, V. P. (2019).** Convolutional neural networks for toxic comment classification. Proceedings of the 10th Hellenic Conference on Artificial Intelligence, SETN 2018, Patras, Greece, pp. 1–6.
10. **Graves, A., Schmidhuber, J. (2005).** Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks, Vol. 18, pp. 602–610.
11. **Hernández-Castañeda, A., Calvo, H., Gelbukh, A., García, F. J. (2017).** Cross-domain deception detection using support vector networks. Soft Computing, Vol. 21, No. 3, pp. 585–595.
12. **Juárez Gambino, O., Calvo, H. (2019).** Predicting emotional reactions to news articles in social networks. Computer Speech & Language, Vol. 58, pp. 280–303.
13. **Kingma, D. P., Ba, J. (2014).** Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.
14. **Kurita, K., Belova, A., Anastasopoulos, A. (2020).** Towards robust toxic content classification. EDSMLS 2020 (The AAAI-20 Workshop on Engineering Dependable and Secure Machine Learning Systems), New York City, United States.
15. **Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., Cambria, E. (2019).** Dialoguernn: An attentive rnn for emotion detection in conversations. Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, Honolulu, Hawaii, USA, pp. 6818–6825.
16. **Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013).** Efficient estimation of word representations in vector space. 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings.
17. **Mustafa, R. U., Ashraf, N., Ahmed, F. S., Ferzund, J., Shahzad, B., Gelbukh, A. (2020).** A multiclass depression detection in social media based on sentiment analysis. 17th International Conference on Information Technology–New Generations (ITNG 2020), Advances in Intelligent Systems and Computing, volume 1134, Las Vegas, Nevada, USA, pp. 659–662.
18. **Ojo, O. E., Gelbukh, A., Calvo, H., Adebanji, O. (2021).** Performance study of n-grams in the analysis of sentiments. Journal of the Nigerian Society of Physical Sciences, pp. 140–143.
19. **Ojo, O. E., Gelbukh, A., Calvo, H., Sidorov, G., Adebanji, O. (2020).** Sentiment analysis in texts on economic domain. Proceedings of the 19th Mexican International Conference on Artificial Intelligence - MICAI2020, Mexico City, Mexico.
20. **Ozoh, P. A., Adigun, A. A., Olayiwola, M. O. (2019).** Identification and classification of toxic comments on social media using machine learning techniques. International Journal of Research and Innovation in Applied Science (IJRIAS), Vol. IV, pp. 142–147.
21. **Pennington, J., Socher, R., Manning, C. (2014).** GloVe: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, pp. 1532–1543.
22. **Poria, S., Cambria, E., Gelbukh, A. (2016).** Aspect extraction for opinion mining with a deep convolutional neural network. Knowledge-Based System, Vol. 108, pp. 42–49.
23. **Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., Stranisci, M. (2018).** An Italian Twitter corpus of hate speech against immigrants. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan.

Article received on 10/12/2021; accepted on 08/03/2022.
Corresponding author is Hiram Calvo.