Detecting Abusive language based on contextual data from Twitter using Machine Learning and Deep Learning Models

Hossain, M. , Haque, S. S. , Aunindita, R. F. , Rakin, A. A. , Aich, A. , Hossain, Md. S. , Rasel, A. A.
School of Data and Sciences
BRAC University
Dhaka, Bangladesh
{mainul.hossain, shataddru.shyan.haque, rudaba.farhin.aunindita, abdullah.al.rakin, ankan.aich, md.sabbir.hossain1}@g.bracu.ac.bd, annajiat@bracu.ac.bd

Abstract

In today's online based world, the use of contextual based symbols has increased drastically. Although online aggression is context-dependent, annotating massive quantities of data is incredibly challenging. Emoji for example, a single emoji can co-occur with various sorts of clearly hostile statements, indicating extralinguistic information. That's why our work focuses on use abusive emojis as a stand-in for acquiring a vocabulary of abusive words. Quality of the study of previous datasets in abusive language detection were inadequate in quantity and quality to train deep learning models efficiently. Online Social network sites such as twitter have a vast array of textual data and such a dataset, " Hate and Abusive Speech on Twitter ", is now available. In this research, we will investigate the impact of several deep learning models on Hate and Abusive Speech on Twitter, discover the possibilities of integrating extra features and context data to train a model for the detection of abusive languages online using contextual data such as emoji. Hate and Abusive Speech on Twitter classifies tweets into 4 labels, "normal", "spam", "hateful" and "abusive". URLs, commonly used emojis and User IDs, are changed as special tokens during data pre-processing. Because hashtags have a significant association with the content of the tweet, we employ hashtag segmentation to extract more information. Tweets are converted into just one encoded vector with 70-character dimensions (26 lower-case alphabets, 10 numbers, and 34 special characters including whitespace). We used both traditional Machine Learning Models as well as Deep Learning Models to compare their respective performances. For Machine Learning models we took Naive Bayes classifier, Support Vector Machines, Logistic Regression and Random Forests. For Deep Learning Models we used Convolutional Neural Networks with ReLU activation and a max pooling layer. We used a Hybrid CNN model as well to evaluate its performance. A Bidirectional RNN is also used as a baseline with cross entropy as the sigmoid function and Adam Optimization. Testing showed that Deep Learning Models are more accurate performance wise with Logistic Regression being the top performer among traditional models. Overall, we conclude that while character level features have improved accuracy of Deep Learning

models, they reduced the overall classification accuracy of the dataset which we conclude is due to the imbalance in data and lack of labels in the data as well.

Keywords

Natural Language Processing, Machine Learning, Deep Learning, Hate speech, Twitter, Emoji's, Contextual Data recognition, Abusive Language Detection.

References:

[1] Corazza, M., Menini, S., Cabrio, E., Tonelli, S., & Villata, S. (2020). Hybrid emoji-based Masked language models for ZERO-SHOT abusive language detection. *Findings of the Association for Computational Linguistics: EMNLP 2020*. https://doi.org/10.18653/v1/2020.findings-emnlp.84

[2] Lee, Y., Yoon, S., & Jung, K. (2018). Comparative studies of Detecting abusive language on Twitter. *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. https://doi.org/10.18653/v1/w18-5113

[3] Pitsilis, G. K., Ramampiaro, H., & Langseth, H. (2018). Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence, 48*(12), 4730–4742. https://doi.org/10.1007/s10489-018-1242-y