Using Attention-based Bidirectional LSTM to Identify Different Categories of Offensive Language Directed Toward Female Celebrities

Sima Sharifirad
Department of computer science,
Halifax, Nova Scotia
s.sharifirad@dal.ca

Stan Matwin
Department of computer science,
Halifax, Nova Scotia
stan@cs.dal.ca

Abstract

Social media posts reflect the emotions, intentions and mental state of the users. Twitter users who harass famous female figures may do so with different intentions and intensities. Recent studies have published datasets focusing on different types of online harassment, vulgar language and emotional intensities. We trained, validate and test our proposed model, attention-based bidirectional neural network, on the three datasets:"online harassment", "vulgar language" and "valance" and achieved state of the art performance in two of the datasets. We report F1 score for each dataset separately along with the final precision, recall and macro-averaged F1 score. In addition, we identify ten female figures from different professions and racial backgrounds who have experienced harassment on Twitter. We tested the trained models on ten collected corpuses each related to one famous female figure to predict the type of harassing language, the type of vulgar language and the degree of intensity of language occurring on their social platforms. Interestingly, the achieved results show different patterns of linguistic use targeting different racial background and occupations. The contribution of this study is two-fold. From the technical perspective, our proposed methodology is shown to be effective with a good margin in comparison to the previous state-of-the-art results on one of the two available datasets. From the social perspective, we introduce a methodology which can unlock facts about the nature of offensive language targeting women on online social platforms. The collected dataset will be shared publicly for further investigation.

1 Introduction

As the internet becomes a human experience, it becomes critical to feel safe online. Often, when talking about violence or harassment against women, online harassment is ignored. In addition, online harassment is not often one dimensional: women from different social groups have been harassed because of their race, their ethnicity, their sexuality, the religious identity and even because of their disabilities (WMC Speech Project, 2018). The goal of our study is to present a comprehensive and multi-faceted experiment on the types of online harassment, vulgar language and intensity of the tweets happening around famous female figures on social media using bidirectional attention-based neural network. To this end, the contribution of the paper as follows:

- Training and testing bidirectional LSTM with an attention layer on three public datasets: "online harassment" (S.Sharifirad and S.Matwin, 2018), "vulgar language" (Holgate et al., 2018) and "valance" (M.Mohammad et al., 2018) and achieving a state-of-the-art results with good margins in two of the datasets
- Using the python Twitter API to collect five hundred tweets directed to each of the ten selected female figures over a period of one month
- Testing the three pre-trained models on each of the ten datasets to understand the online harassment types, the vulgar language types and the intensity of emotions
- Demonstrating the classification results of the proposed model directed to famous female figures on ten datasets and addressing the impact of occupation and racial background on the results

2 Experiment

In this paper, we used bidirectional RNN because of its unique ability to memorize the content of a word both in future and past. We accompanied this model with an attention layer to learn a weight for each word. After extracting the textual features, the features were fed into a softmax classifier for classification. We used drop out and consider Adam as the optimization algorithm. We used a grid search to determine the best parameters for dimensions of hidden layers from {500, 1000, 2000}, drop out ratio of 0.5, learning rate from {1, 0.1, 0.01, 0.001} for each dataset. The model was implemented in Keras inspired by Peng Zhou (2016). The first experiment is related to detecting different types of online harassment. This dataset was proposed by Sharifirad and Matwin (2018) and contains three categories of "indirect harassment", "physical harassment" and "sexual harassment". This dataset is quite imbalanced, we first balanced the dataset using Synthetic Minority Oversampling Technique (SMOTE). After fine tuning our model, the F1 score for the three categories is as follows: Indirect harassment-89.8, Physical harassment-92.1 and sexual harassment-92.8. The macro averaged F1-score is 91.5. Even though, in this dataset, we didn't have significant improvement, the F1 measure was good enough to test on our collected dataset.The second dataset is related to detecting six types of vulgar language (Holgate et al., 2018). We split each of the three datasets into train, validation and test set. We report the final precision, recall and F1 score and report F1-score for each class separately. Our predictive model F1 score for each of the vulgar classes is as follows: Aggression-85.6, Emotion-82.3, Emphasize-86.1, Signal Group Identity-80.5, Auxiliary-84.7, Not vulgar-88.1. Our model achieves a macro averaged F1 score of 84.5 across the six classes. It is a big improvement in comparison to the baseline set by Holgate et al., (2018). The third dataset is composed of different levels of valence or intensity of emotion in the tweets. This dataset was released in SemEval 2018, "Affect classification" (M.Mohammad et al,, 2018). After deploying and fine tuning our model, we had superior performance, 90.2 in accuracy in comparison to previous state-of-the art which was 83.6.In this study, we consider ten female figures of different racial backgrounds and professions who have experienced online harassment. We collected five hundred tweets on the duration of one month using the python Twitter API. These tweets have been released in author GitHub for further investigation. After training, validating and testing three separate models on three datasets, we tested these three models on ten collected datasets each around one female figure. Looking into the table of labels related to famous figures revealed very interesting findings for us.Generally, black female figures on average receive more aggressive tweets with high negative valence. In addition, they receive more of the tweets which show signal group identity which means some users write tweets which focused on the black community and try to provoke other members in their group to write toxic comments. In addition, they have more tweets in the category of auxiliary vulgar language in comparison to other racial backgrounds. White female figures also received aggressive and emotional tweets. However, there were a few tweets about signal group identity and auxiliary. Female athletes irrespective of their racial background receive a high number of physical harassment, rape and death threat tweets. Interestingly enough, they receive indirect harassment as well which shows there are users who harass female athletes by indirectly questioning their way of playing or their competency in sports in comparison to their male colleagues. Those female figures who are actresses, they receive two main types of harassment, physical and sexual harassment.Usually, In social media they are not get compared to their male peers but are harassed , are threatened with death or rape or are received tweets which view them as sex objects.Those female figures who are activists or are civil right speakers receive tweets with a lot of aggression, negative emotions and a high valence of physical or sexual harassment. If they receive tweets which were moderately negative, they are those tweets which invite them to be silent or condemn their speech but these tweets don't contain any swear words. These findings are not only in line with Delisle et al. (2018) research but also provides more information on different occupations.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503–512.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

WMC Speech Project. 2018. *Online Abuse 101*. http://www.womensmediacenter.com/speech-project/online-abuse-101/.

S.Sharifirad and S.Matwin. 2018. *Different types of sexist language on twitter and the gender footprint.*. CICLing.

olgate, E.; Cachola, I.; Preo tiuc-Pietro, D.; and L. 2018. *Why swear? Analyzing and inferring the intentions of vulgar expressions.*. EMNLP, 4405-4414.

Mohammad, Saif M. and Bravo-Marquez, Felipe and Salameh, Mohammad and Kiritchenko, Svetlana. 2018. *SemEval-2018 Task 1: Affect in Tweets*. Proceedings of International Workshop on Semantic Evaluation (SemEval-2018).

Laure Delisle, Alfredo Kalaitzis, Krzysztof Majewski, Archy de Berker, Milena Marin and Julien Cornebise. 2018. *A large-scale crowd-sourced analysis of abuse against women journalists and politicians on Twitter*. NIPS.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi , Bingchen Li, Hongwei Hao, Bo Xu. 2016. *Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification*. ACL.