Team 10 : Hate Speech Detection

Using Hurtlex Lexicon

Miran M. Rashid   Salim Fares
supervisor : Dr. Jelena Mitrovic

July 2020

1 Related Work

Recent years have seen an increasing number of research on hate speech detection as well as other. But there has been limited literature on the problem of Hate Speech detection on social media. Albadi et al. (2018) they provided a first attempt to investigate the problem of religious hate speech detection in Arabic Twitter. Their GRU-based RNN with pre-trained word embeddings model gave the best performance with 0.79 accuracy, 076 precision, 0.78 recall, 0.77 F1-Score [1]. It is quite challenging to have an accurate classifier to detect hate speech since offensive language which doesn't meet the definition of hate speech causes misclassification. Any supervised approach causes misclassifying hate speech with offensive language (Burnap and Williams 2015)[2]. (Thomas Davidson et al, 2017)by using multi-class classifier with 5-fold cross-validation to predict HS, offensive language and neither on 25K dataset by Logistic regression L1 penalty to select the best feature then L2 regularization to predict the class, the model obtains F1 score 0.91[3]. (Chowdhury et al., 2019) their methodology has successfully demonstrated how community interaction can be leveraged to tackle downstream NLP tasks like detection of religious hate speech. They obtained a highest f1-score of 0.78. This is achieved using LSTM + CNN + CISNet model. And their LSTM + CNN + NODE2VEC (ARHNet) model achieved the best recall 0.89, while GRU-NODE2VEC demonstrates the highest precision 0.74, and the Bi-GRU-CNN-NODE2VEC model achieves the highest accuracy 0.81[4]. (Ousidhoum et al., 2019) they presented a multilingual hate speech dataset of English, French, and Arabic tweets. They performed multilingual and multitask learning on their corpora and showed that deep learning models perform better than traditional BOW-based models in most of the multilabel classification tasks. Their best results for Arabic according to their experiments: Best model for Directness is STSL model (Single Task Single Language) with 0.84 F1-Score. Best model for Hostility Type are STML model (Single Task Multi Language) and MTML model (Multi Task Multi Languages) with 0.35 F1-Score. Best model for Target Attribute is STSL model with 0.63 F1-Score. Best model for Target Group is LR (Logistic Regression) with 0.40 F1-Score. Best model for Annotator's Sentiment is MTML model with 0.21 F1-Score[5].

References

[1] N. Albadi, M. Kurdi, and S. Mishra. "Are they Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere". In: *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 2018, pp. 69–76.

[2] Pete Burnap and Matthew L. Williams. "Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making". In: *Policy & Internet* 7.2 (2015), pp. 223–242. DOI: 10.1002/poi3.85. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/poi3.85. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.85.

[3] Thomas Davidson et al. "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, 2017, pp. 512–515.

[4] Arijit Ghosh Chowdhury et al. "ARHNet - Leveraging Community Interaction for Detection of Religious Hate Speech in Arabic". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 273–280. DOI: 10.18653/v1/P19-2038. URL: https://www.aclweb.org/anthology/P19-2038.

[5] Nedjma Ousidhoum et al. "Multilingual and Multi-Aspect Hate Speech Analysis". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4675–4684. DOI: 10.18653/v1/D19-1474. URL: https://www.aclweb.org/anthology/D19-1474.