A comparative study of different features for efficient automatic hate speech detection

Nicolas Zampieri, Irina Illina, Dominique Fohr

A comparative study of different features for efficient automatic hate speech detection

Nicolas Zampieri, Irina Illina, Dominique Fohr
Universite de Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France
{nicolas.zampieri, irina.illina, dominique.fohr}@loria.fr

Abstract

Commonly, Hate Speech (HS) is defined as any communication that disparages a person or a group on the basis of some characteristic (race, colour, ethnicity, gender, sexual orientation, nationality, etc. (Nockeby, 2000)). Due to the massive activities of user-generator on social networks (around 500 million tweets per day) Hate Speech is continuously increasing on the web.

Recent initiatives, such as SemEval2019 shared task 5 Hateval2019 (Basile et al., 2019) contribute to the development of automatic hate speech detection systems (HSD) by making available annotated hateful corpus. We focus our research on automatic classification of hateful tweets, which are the first sub-task of Hateval2019. The best Hateval2019 HSD system was FERMI (Indurthi et al., 2019) with 65.1 % macro-F1 score on the test corpus. This system used sentence embeddings, Universal Sentence Encoder (USE) (Cer et al., 2018) as input of a Support Vector Machine classifier.

In this article, we study the impact of different features on an HSD system. We use deep neural network (DNN) based classifier with USE. We investigate the word level features, such as lexicon of hateful words (HFW), Part of Speech (POS), uppercase letters (UP), punctuation marks (PUNCT), the ratio of the number of times a word appears in hateful tweets compared to the total number of times that word appears (RatioHW) ; and the emojis (EMO). We think that these features are relevant because they carry feelings. For instance, cases (UP) and punctuations (PUNCT) can carry the intonation of the tweets and can be used to express a hateful content. For HFW features, we tag each word of tweets as hateful or not using the Hatebase lexicon (*Hatebase.org*) and we associate a binary value to each word. For POS features, we use twpipe (Liu et al., 2018) for tagging the words and this information is coded as an one-hot vector. For emojis, we generate an embedding vector using emoji2vec tools (Eisner et al., 2016). The input of our neural network consists of the USE vector and our additional features. We used convolutional neural networks (CNN) as binary classifier. We performed the experiments on the HateEval2019 corpus to study the influence of each proposed feature. Our baseline system without proposed features achieves 65.7% of macro-F1 score on the test corpus. Surprisingly, HFW degrades the system performance and decreases the macro-F1 by 14 points compared to the baseline. This can be due to the fact that some words are hateful only in a particular context. UP, RatioHW and PUNCT slightly degrade the baseline system. The POS features do not change the baseline system result and so are probably not correlated to the hate speech. The best result is obtained using EMO features with 66.0% of macro-F1. EMOs are largely used to transmit emotions. In our system,they are modeled by a specific embedding vector. USE does not take into account the emojis. Therefore, EMOs give additional information to USE about the hateful content of tweets.

This work was performed in the context of the Franco-German ANR projet M-PHASIS.

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5 : Multilingual detection of hate speech

against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pages 169–174, Brussels, Belgium, November. Association for Computational Linguistics.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec : Learning emoji representations from their description. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 48–54, Austin, TX, USA, November. Association for Computational Linguistics.

Vijayasaradhi Indurthi, Bakhtiyar Syed, Manish Shrivastava, Nikhil Chakravartula, Manish Gupta, and Vasudeva Varma. 2019. FERMI at SemEval-2019 task 5 : Using sentence embeddings to identify hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 70–74, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A. Smith. 2018. Parsing tweets into Universal Dependencies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, pages 965–975, New Orleans, Louisiana, June. Association for Computational Linguistics.

John T. Nockeby. 2000. Hate speech. In Leonard W. Levy, Kenneth L. Karst, and Dennis J. Mahoney, editors, *Encyclopedia of the American Constitution*, pages 1277–1279. Macmillan 2nd edition.