MACHINE LEARNING ALGORITHMS FOR HATE SPEECH DETECTION

126003131 Keshav K
126003209 Raahul Kumar N
126156144 Shruti Rajesh

ABSTRACT:

Over the last two decades the proliferation of technology has enabled humans to connect with each other better through social media platforms and people exercise their freedom of speech to the fullest. With the increased platforms to share their opinions, hate speech has also flourished. Usually propagated under the guise of anonymity, hate speech does detrimental damage to anyone involved explicitly in the exchange and implicitly coming across the interaction. The primary aim of this paper is to review Machine Learning algorithms and techniques for detecting such hate speech in social media. In this study, we will examine the baseline components for hate speech classification - data collection of a published or custom dataset and its subsequent exploration, feature extraction, feature selection by reduction of dimensions, hate speech classifier selection and training, and model evaluation. In this literature, we plan on proposing different ensemble approaches using classical Machine Learning methods, Deep Learning techniques and dimensionality reduction algorithms; and we would also use different performance metrics to evaluate our approaches such as Precision, F1 scores, Accuracy, and Recall parameters. The results from these will be assessed and compared through learning curves to have an estimation of what the best model among the many approaches used will be, and the drawbacks and liabilities of each model will be discussed extensively alongside their performance parameters to have a better sense of optimizing the model or opting to stratify using a better algorithm. This study will have 4 major contributions. Firstly, it would equip the readers with all the critical steps involved in hate speech detection. Secondly, it would give a thorough analysis on all the models involved. Third, it would include any gaps in research and challenges that will be identified in the process. Lastly we will take a closer look at scalability of the datasets used and criticize the multicultural impact that cyber hate has, throughout the world. The entire domain of hate speech involves recurrent updating of the datasets as vocabulary and modern terms are ever-changing, which will also be briefly elucidated. This study and implementation can immensely help learners, researchers, the public and professionals alike.

Keywords: Hate speech classification, Data collection, Feature extraction, Dimensionality reduction, Ensemble approaches, Deep Learning techniques, Performance metrics (Precision, F1 scores, Accuracy, Recall).

Guide: Dr. Senthil Selvan N