Detecting Hate Speech And Cyberbullying Using Natural Language Processing

Authors : Nádia Pereira, Paula Ferreira, Sofia Francisco, Sofia Oliveira, Sidclay Souza, Paula Paulino, Ana Margarida Veiga Simão

Abstract : Social media has progressed into a platform for hate speech among its users, and thus, there is an increasing need to develop automatic detection classifiers of offense and conflicts to help decrease the prevalence of such incidents. Online communication can be used to intentionally harm someone, which is why such classifiers could be essential in social networks. A possible application of these classifiers is the automatic detection of cyberbullying. Even though identifying the aggressive language used in online interactions could be important to build cyberbullying datasets, there are other criteria that must be considered. Being able to capture the language, which is indicative of the intent to harm others in a specific context of online interaction is fundamental. Offense and hate speech may be the foundation of online conflicts, which have become commonly used in social media and are an emergent research focus in machine learning and natural language processing. This study presents two Portuguese language offense-related datasets which serve as examples for future research and extend the study of the topic. The first is similar to other offense detection related datasets and is entitled Aggressiveness dataset. The second is a novelty because of the use of the history of the interaction between users and is entitled the Conflicts/Attacks dataset. Both datasets were developed in different phases. Firstly, we performed a content analysis of verbal aggression witnessed by adolescents in situations of cyberbullying. Secondly, we computed frequency analyses from the previous phase to gather lexical and linguistic cues used to identify potentially aggressive conflicts and attacks which were posted on Twitter. Thirdly, thorough annotation of real tweets was performed byindependent postgraduate educational psychologists with experience in cyberbullying research. Lastly, we benchmarked these datasets with other machine learning classifiers.

Keywords : aggression, classifiers, cyberbullying, datasets, hate speech, machine learning
Conference Title : ICDID 2022 : International Conference on Digital Innovation and Development
Conference Location : Lisbon, Portugal
Conference Dates : September 20-21, 2022

Open Science Index, Computer and Information Engineering Vol:16, No:09, 2022 publications.waset.org/abstracts/142382.pdf