Cyber-bullying and Cyber-harassment Detection using Supervised and Unsupervised Machine Learning Techniques in Arabic Social Media Contents

By

Amal Mohmmad Aldajah

Supervisor

Dr. Tarek G Kanan

Abstract

This study aims to propose a system to detect Cyber-Bullying and Cyber-Harassment posts from Twitter and Facebook by using Machine Learning algorithms; Classification and Clustering. The great turnout of using social media in recent years such as Facebook and Twitter has led to the emergence of many negative behaviors and immoral phenomena, including Cyber-bullying and Cyber-harassment. The widespread of social media among the users has led to increased interaction and communication between them electronically. It also allows users to publish their written and multimedia content in addition to the ability to express feelings and emotions about a particular subject. Many research exploits the spread and communication between users and interacts with each other in topics raised through social media to discover emotions, identify user opinions, and discover behaviors and negative phenomena of users through using Machine Learning algorithms.

It is found that these behaviors are dangerous activities and have negative psychological, health and social effects on the user. There are many studies that detect these phenomena but these studies are limited to the English language, we focus through our research on the Arabic language. Despite the difficulty of the Arabic language as it is a derivative language with high morphology. We were able to address this language and simplify it through the use of Arabic Natural Language Processing. Three Arabic natural language processing tools have been implemented to help with our work; Stop word

Removal, Normalization, and Stemming. Despite the increasing prevalence of these negative phenomena through the means of social communication, there is a significant lack of awareness of the effects and disadvantages of these phenomena and lack of knowledge of the penalties of cyber-bullying and cyber-harassment, so we have created pages on Facebook and Twitter to spread the knowledge of cyber-bullying and cyber-harassment and identify their psychological, health and social effects. In addition, we opened a channel on YouTube that publishes several awareness videos. We collected data from Facebook and Twitter which included 6,138 post Twitter and Facebook. Netvizz Application was used to collect datasets from Facebook and the RStudio Software to collect data from Twitter. We applied five classification algorithms to the dataset including; (K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naive Bayes (NB), Random forests (RF), and Decision Tree (J48)). To calculate the performance of the algorithms we calculated F-Measure, Recall, and Precision. We applied two clustering algorithms to the dataset including; K-Means and Expectation- Maximization (EM). To calculate the performance of the algorithms we calculate Training Time and Cluster Sum of Squared Error.

It is concluded that the Random Forest algorithm yields the highest values of F-Measure (94.70%) followed by SVM, NB, J48, and KNN respectively when algorithms are applied to all datasets and with applied pre-processing; Normalization, Stop-Word Removal, and Stemming. The same results occur when applied all datasets without stemming and without stop-word removal, Random Forest gives the highest value. SVM gives higher F-Measure values (94.40%) when separating datasets into Facebook Posts and Twitter Tweets then applying algorithms to them