Using Sentiment Analysis to Identify Bullying Using Twitter

Hoanh Nguyen
University Massachusetts Lowell
1 University Ave
Lowell, MA 01854, USA
soujiroboi@gmail.com

Michael Meding
University Massachusetts Lowell
1 University Ave
Lowell, MA 01854, USA
mikeymeding@gmail.com

Abstract

Twitter is a social network where users can communicate publicly with short text statements called tweets. It should come as no surprise that not all of this communication has good intention. With the rise of social media in youth, Twitter has gone from a calm social environment to a hostile place to connect with others This program identifies hostile users who have a high likelihood of being bullies through sentiment analysis of all elements of their tweets. We did this by taking data from twitter and analysing the aggressiveness of given tweets and hashtags using the SentiWordNet database and improved using Harvard Inquirer. This is then put into a Machine Learning algorithm to determine if a user can be classified as a bully based on current and past information.

1 Introduction

This program identifies Twitter users who are bullies to those around them. To accomplish this task the program trains a binary classifier to decide if a single tweet is bullying or not. However, later research has shown that bullying cannot fully be defined by a simple binary ”yes” or ”no”. Bullying in a broad sense can be classified into many different categories beyond that of just the bully and the victim. When observing bullying in action on twitter you will notice that there are many supporting people who do not have anything to do directly with a given bullying situation. This leads to a large amount of ambiguity when trying to decern whether someone is a bully, defender, victim, or accuser.

For something to be cleanly identified by a computer as being bullying is to have both an aggressive statement and a subject or direction of that anger. Often times the subject of an aggressive tweet is implied making it very difficult for a machine to find a subject. The solution to this was an assisted machine learning algorithm using a brute force approach which has a high accuracy when tweets directly mention a subject. Given well annotated data the results were surprisingly good given small data sets. The data sets were manually annotated by a human who could only use the information contained within the tweet to decide if that user is a bully or not.

2 Related Work

The University of Wisconsin did a project two years ago on the study of bullying in twitter that heralded lots of media coverage including several articles from Huffington Post and Time Magazine. The first version of their code was made public that same year and suffice to say that this project has improved on their initial algorithm. Their code used a set of static words as search terms for tweets then classified by a very small existing bullying model. This model did yield some results but the efficacy left much to be desired.

3 Methodology

3.1 Machine Learning Methods

3.2 Data Sets

Overview The data for this project was gathered by a data mining algorithm. The algorithm uses a lexicon of twitter hashtags and search terms that have been tagged with a sentiment score. It goes through this lexicon to find the words which have medium to high negative sentiment score. These words are then searched for individually and the top results are then stored for analysis. These results are filtered for spam and for relevancy before being saved. Based on the sentiment score of the search term it is able to control the size of the output file as individual search terms are limited to 40 tweets per query. From there tweets are manually tagged bullying or not_bullying to create a gold standard for the given data.

Lexicons This project included several lexicons which were used to develop features. The unigram hashtag sentiment lexicon was used for the generation of aggressive search terms to get the best potential for bullying tweets. The lexicon was organized as ¡term¿tab¡sentiment score¿tab¡positive count¿tab¡negative count¿. This lexicon was compiled over a large amount of twitter data to give good scores. For our project we only cared about the terms which related to negative sentiment with a relative sentiment score below -4.999. Another useful lexicon for this project was the emoticon lexicon. This included emoticons such as :D and :(and their related sentiment in 3 categories, positive, negative, and neutral. Each of these categories was used as a feature in the identification of a bully.

Corpus The corpus for this project blah blah blah....

3.3 Evaluation

4 Results

5 Future Improvements

Acknowledgments

We thank Alison Rose, and Uwe Meding for their contributions in annotating our data set for this project. Without their efforts this project would not have been possible. Additionally, we would like to thank Willam Boag for providing us with lexicons from a similar project which proved to be immensely helpful.