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. However, it should come as no surprise that not all of these tweets have good intention. With the rise of social media usage in youth, Twitter has gone from a calm social environment to connect with others to a hostile place. This program identifies users who have a high likelihood of being a bully through the use of sentiment analysis in all of their tweets. We did this by taking data from Twitter and analysing the aggressiveness of a tweets and hashtags using both the SentiWord-Net database and Harvard Inquirer. This data is then run through a Machine Learning algorithm to determine if a user can be classified as a bully based on similar tweets.

1 Introduction

What is bullying? Bullying is when a person is constantly being exposed to negative actions by another individual or group of individuals and can involve both physical and mental abuse. Some common forms of mental bullying include name calling, making threats, and spreading rumours. Since the creation of social media sites, bullying has become a problem outside of the school environment and is now considered a serious national health issue among adolescents [0]. Bullies are using the internet as an outlet to post negative comments about someone, and unlike in a school environment, they rarely see backlash for what they post on social media sites.

This program identifies Twitter users who are bulling those around them. To accomplish this task, the program trains a binary classifier to decide if a single tweet is bullying or not. However, further 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. For example, when observing a bullying in action on Twitter, you will notice that there are a lot of supporting users who do not have anything to do directly with a given bullying situation. This leads to a large amount of ambiguity when trying to decide whether someone is a bully, defender, victim, or accuser.

For a tweet to be cleanly identified by a computer as being bullying it must have both an aggressive statement and a subject or direction towards that anger. One of the major flaws with this is that often times the subject of an aggressive tweet is implied, thus making it very difficult for a machine to identify a subject. Our solution to this problem was to use an assisted machine learning algorithm. The data sets were manually annotated by a human who could only use the information contained within a given tweet to decide if that user is a bully or not. By using a brute force approach, we achieved a high accuracy when tweets directly mention a subject. Given the small data set, the annotated results were surprising.

2 Related Work

Two years ago, the University of Wisconsin completed a project on the study of bullying in Twit-

ter that heralded lots of media coverage, including several articles from *The Huffington Post* and *Time Magazine*. The University of Wisconsin's code used a set of static words as search terms for tweets, then classified them based on a very small pre-existing bullying model. This model did yield some results but the efficacy left much to be desired. The first version of the University's code was made public that same year the project was released and our project has improved on their initial algorithm.

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 the 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. Then the results are filtered for spam and relevancy before being saved. Based on the sentiment score of the search term, the lexicon is able to control the size of the output file because individual search terms are limited to 40 tweets per query. From there, the tweets are manually tagged bullying or not_bullying to create a standard for future 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 in order to achieve the best potential for bullying tweets. The lexicon was organized into four different categories: term, sentiment score, positive count, and negative count. The 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 After going through the lexicon of negative sentiment words we end up with an array of negative words. This array now represents our search terms. Using the Twitter API, the program queries the top 40 English tweets one at a time for each word in the array of negative words. We are limited to a maximum of 40 tweets per request and 180 requests every 15 minutes. After each request has finished, the received tweets are pushed down and appended into an untagged corpus text file which is then ready for the final stage of corpus construction.

Annotation The final stage of building the data set for this project is manual tagging. After exhausting the array of negative words, we end up with a large file of tweets which has a high potential of being tagged bullying. As this program uses an assisted machine learning algorithm, it requires a basis of comparison to verify the tagging efficacy. Therefore, it is required for someone to manually go through each tweet and tag each as either bullying or not_bullying. This, as you can imagine, is exceptionally time consuming and has required the combined efforts of Alison Rose, Uwe Meding, and Michael Meding.

3.3 Evaluation

4 Results

5 Future Improvements

Acknowledgments

References

Jun-Ming Xu, Xiaojin Zhu, Amy Bellmore. Fast learning for sentiment analysis on bullying.