Detecting threats of violence in YouTube comments

Anonymous NAACL submission

Abstract

Here comes an abstract for this article

1 Introduction

Threats of violence is an increasingly common occurrence in online discussions. It disproportionally affects women and minorities, often to the point of effectively eliminating them from taking part in discussions online. Moderators of social networks operate on such a large scale that manually reading all posts is an insurmountable task. Methods for automatically detecting threats could therefore potentially be very helpful, both to moderators of social networks, and to the members of those networks.

In this article, we evaluate different types of features for the task of detecting threats of violence in YouTube comments. We show that ...

2 Previous work

There is little previous work specifically devoted to the detection of threats of violence in text, however, there is previous work which examines other types of closely related phenomena, such as cyberbullying and hate-speech.

Dinakar et al. (2011) proposes a method for the detection of cyberbullying by targeting combinations of profane or negative words, and words related to several predetermined sensitive topics. Their data set consists of over 50,000 YouTube comments taken from videos about controversial topics. The topics included sexuality, race, culture and intelligence. The comments were grouped by topic, and then 12 % of these were manually annotated to check that they were placed in the right category.

The first stage of the detection was the same across all categories. It consisted of using a lexicon of negative words and a list of profane words, as well as part-of-speech tags from the training data that were correlated with bullying. The second stage was category-specific, and used commonly observed uni- and bigrams from each category as features. The experiments reported accuracies from 63 % to 80 %, but did not report precision or recall.

3 The YouTube threat data set

The YouTube threat data set is comprised of user-written comments from eight different YouTube videos (Hammer, 2014). A comment consists of a set of sentences, each of them manually annotated to be either a threat of violence (or support for a threat of violence) or not. The data set furthermore records the username[1] of the user that posted the comment. The eight videos that the comments were posted to cover religious and political topics like halal slaughter, immigration, Anders Behring Breivik, Jihad, etc. (Hammer, 2014).

The data set consists of 9,845 comments, comprised of 28,643 sentences, see table 1. In total there are 402,673 tokens in the sentences in the data set. There are 1,285 comments containing threats, and 1,384 sentences containing threats, as seen in table

[1] In 2013, YouTube changed its commenting system from using unique usernames, to using ”real names”, like Facebook and other sites (YouTube, 2013). Some accounts, however, did not provide real names, so they continue to only be identified by their usernames.

	Commments	Sentences	Users posting
Total	9,845	28,643	5,483
Threats	1,285	1,384	992

Table 1: Number of comments, sentences and users in the YouTube threat data set

1. (Hammer, 2014) report inter annotator agreement on this data set to be 98 %, as calculated on 120 of the comments, doubly annotated for evaluation.

Figure 1 provides some examples of comments containing threats of violence taken from the data set. The first line is the username or name, and the subsequent lines are the sentences of the comment. An empty line indicates the end of a comment. The sentences are annotated with a number indicating whether they contain a threat of violence (1), or not (0).

```
timpa666
1    and i will kill every fucking muslim and arab!

NimsXdimensions
0    Need a solution?
1    Drop one good ol' nuke on that black toilet in Mecca.

Ammar Alozaibi
1    Funny, We will conquer you all in just few years, U will be
     my slave and your women will be my Sex Toy in Bed.
```

Figure 1: Examples of comments from the data set.

4 Experiments

4.1 Experimental setup

4.2 Results

5 Discussion

References

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *Proceedings of The Social Mobile Web*.

Hugo Lewi Hammer. 2014. Detecting threats of violence in online discussion using bigrams of important words.

YouTube. 2013. We hear you: Better commenting coming to YouTube. youtube-global.blogspot.no/2013/09/youtube-new-comments.html/. [Online; accessed May 12th 2015].