Detecting threats of violence in online discussions

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 big 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 ...

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.

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.

[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

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 the YouTube threat data setconsists 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. (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 contains examples of some 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.

LegitZombieSlayer1
0     As long as i'm alive you'll have no victory.
1     I'll kill all you cunts
```

Figure 1: Examples of comments from the data set.

4 Experiments

4.1 Experimental setup

4.2 Results

5 Discussion

References

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].