Detecting threats of violence in online discussions using bigrams of important words

Hugo Lewi Hammer
Oslo and Akershus University College of Applied Sciences
Department of Computer Science
N-0166 Oslo, Norway
Email: hugo.hammer@hioa.no

***Abstract*—Making violent threats towards minorities like immigrants or homosexuals is increasingly common on the Internet. We present a method to automatically detect threats of violence using machine learning. A material of 24,840 sentences from YouTube was manually annotated as violent threats or not, and was used to train and test the machine learning model. Detecting threats of violence works quit well with an error of classifying a violent sentence as not violent of about 10% when the error of classifying a non-violent sentence as violent is adjusted to 5%. The best classification performance is achieved by including features that combine specially chosen important words and the distance between those in the sentence.**

I. Introduction

Over the past years there has been an alarming growth in hate against minorities like Muslims, Jews, Gypsies and gays on the Internet [1], and experts are concerned that individuals influenced by this web content may resort to violence as a result [2], [3].

The main aim of this paper is to evaluate the potential of using different machine learning approaches to detect sentences in hateful online discussions that contain a threat of or sympathy with violence (for short just called threats of violence in the rest of the article).

II. Sentence features to detect threats of violence

Most classification methods within text mining are based on the so called document term matrix, also referred to as bag-of-words or unigram.

We expect that a threat of violence often should contain the subject that wants to perform the violence, like 'I' or 'we', some aggressive words like 'kill', 'bomb', 'nuke', 'gun', etc, as well as the target for the violence, like 'Muslims', 'Jews', 'women', 'bastards', 'sandniggers' and so on. Potentially important features from the sentences therefore are bigrams of such important words.

Naturally we expect that a combination of important words like 'I-kill' is more important if 'I' and 'kill' are close to each other in the sentence, because then it is more likely that 'I' is related to 'kill'. For the sentences

Sentence 1: "I will kill Muslims and I will kill Jews"

Sentence 2: "We love to kill Muslims"

The feature matrix becomes E.g. 'we-kill' occurs once in 'sentence 2' with two words between, such that the computation becomes $1/(2+1) = 1/3$.

TABLE I. Feature matrix for the bigram of important words using weight function

	I-kill	kill-Muslims	Muslims-I	kill-Jews	we-kill
sentence 1	1	1	1/2	1	0
sentence 2	0	1	0	0	1/3

The selections of features above is based on using a set of important words. We chose those words that were significantly correlated with the response (violent/not-violent sentence). Classification is performed using LASSO logistic regression [4].

III. Evaluation

The text material consisted of all comments on eight YouTube videos, all related to religious or political topics like halal slaughter, immigration, Anders Behring Breivik, Jihad etc. The material consisted of 24,840 sentences with 1,469 threats of violence.

The word distance features performs significantly better then traditional unigram with a rate of wrongly classifying a sentence as non-violent is below 10% compared to 14% for unigram. The rate of wrongly classifying a sentence as violent were adjusted to 5%.

IV. Closing remarks

In this article we have shown how text mining and machine learning can be used to detect threats of or sympathies with violence in online discussions.

References

[1] J. Bartlett, J. Birdwell, and M. Littler, "The rise of populism in Europe can be traced through online behaviour..." Demos, http://www.demos.co.uk/files/Demos_OSIPOP_Book-web_03.pdf?1320601634, 2013, [Online; accessed 21-January-2014].

[2] Ø. Strømmen, *The Dark Net. On Right-Wing Extremism, Counter-Jihadism and Terror in Europe*. Oslo, Norway: Cappelen Damm, 2012.

[3] I. M. Sunde, "Preventing radicalization and violent extremism on the Internet (Norwegian)," The Norwegian Police University College 2013:1, 2013.

[4] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010. [Online]. Available: http://www.jstatsoft.org/v33/i01/

978-1-4799-6364-5/14 $31.00 © 2014 IEEE
DOI 10.1109/JISIC.2014.64