Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 236 (2016) 71 – 75

Procedia
Social and Behavioral Sciences

International Conference on Communication in Multicultural Society, CMSC 2015, 6-8 December 2015, Moscow, Russian Federation

Automatic detection of verbal aggression for Russian and American imageboards

Denis Gordeev[a,b,*]

[a]*National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Kashirskoe shosse 31, Moscow 115409, Russian Federation*
[b]*Moscow State Linguistic University, Ostozhenka, 38, Moscow 119034, Russian Federation*

Abstract

The problem of aggression for Internet communities is rampant. Anonymous forums usually called imageboards are notorious for their aggressive and deviant behaviour even in comparison with other Internet communities. This study is aimed at learning ways of automatic detection of verbal aggression for the most popular American (4chan.org) and Russian (2ch.hk) imageboards. The study material consists of 1,802,789 messages. The machine learning algorithm word2vec was applied to detect the state of aggression. A decent result is obtained for English (88%), the results for Russian are yet to be improved.

Keywords: Aggression; word2vec; imageboard; 4chan; 2ch; cyberbullying; random forest

1. Introduction

The Internet is sometimes considered a quite violent and rude place. Many people, especially active users, face with cyberbullying and other expressions of aggression on a daily basis. For example, the U.S. Department of Health & Human Services has launched an initiative to stop bullying, including Internet bullying [A]. According to the article 282 from the Russian criminal code, hate speech on the Internet is punishable by a fine of up to 300 thousand rubles or a sentence of up to 4 years [B]. However, this law does not give any criteria for distinguishing messages

* Corresponding author. Tel.: +7-495-788-5699; fax: +7-499-324-2111.
E-mail address: DIGordeyev@mephi.ru

doi:10.1016/j.sbspro.2016.12.022

arousing hate and it is a task for linguists. Imageboards that have been a buzzword for a while are considered a truly epicentre of all kind of unruly behaviours that we can find on the Net. For example, they are called 'the Internet hate machine' (Bernstein, Monroy-Hernández, Harry, André, Panovich, and Vargas, 2011).

Imageboards are usual Internet forums with no registration. Messages contain no personal details, only the text, date and email. However, registration mechanisms are not implemented and emails are not checked. Personal tripcodes are the only means to state your identity but they are used only in about 4% of cases (Bernstein, Monroy-Hernández, Harry, André, Panovich, and Vargas, 2011). It is only natural that aggression will flourish in such an environment where nobody can track you and where there are no social limits. Nevertheless, Potapova and Gordeev (2015) have shown that it may be not true for Russian Internet communities, although the results are still disputed.

In this research, we study aggression in the environment where it is vividly presented and is not constrained by social boundaries. This research is also important because it is one of the first works on automatic detection of verbal aggression. We also publish our trained neural model online that can be used by other scientists to find word similarities for imageboards and compare them with other sites. Moreover, our methods may be used for training other word similarities models, but the procedure may change for languages that have no explicit word boundaries and in other difficult cases.

2. Related works

Many researchers deal with aggression and its representation on the Internet. Potapova has been investigating aggression (Potapova and Komalova, 2014) and compiled a Russian dictionary containing words describing this emotional state (Potapova and Komalova, 2015). Bernstein has conducted a research on 4chan and imageboard culture (Bernstein, Monroy-Hernández, Harry, André, Panovich, and Vargas, 2011). The task of sentiment analysis is rather close to aggression analysis because both deal with detection of different human emotions. Twitter and social networks sentiment analysis is especially close to our research field, because the majority of anonymous forums messages are short, e.g. a 4chan message contains 15 words on average (Potapova and Gordeev, 2015) and there are no more than 140 symbols for a Twitter post. Numerous papers has been published on this and adjacent topics in recent years. Cerrea et al. studied the influence of complete anonymity on the users' behavior (Correa, Silva, Mondal, Benevenuto, and Gummadi, 2015) in comparison with partial anonymity of Twitter. They have found that users tend to be more open and are more ready to express negative emotions (not only aggression) in anonymous environment. However, they have studied a site Whisper designed to share secrets and confessions, and it may influence their results. Martínez-Cámara has conducted an overview of different methods for Twitter sentiment analysis (Martínez-Cámara, Martín-Valdivia, Ureña-López, and Montejo-Ráez, 2014). Another research was done by Dos Santos. He successfully (from 76% to 88% for various measurement sets) detected the sentiment for Twitter messages (Dos Santos, 2014) without using any handcrafted features. Tang and Wei analyzed Twitter sentiments using emoticons, smileys and neural networks (Tang, Wei, Yang, Zhou, Liu, and Qin, 2014). As we see, many modern studies use machine learning and neural networks methods for sentiment detection. However, Paltoglou (2012) asserts that 'unsupervised' dictionary-based methods outperform 'state-of-the-art' machine learning. Nevertheless, he does not mention any deep learning or neural network-based algorithms, and his results are difficult to apply to other languages, besides English.

3. Methods and materials

Our study is focused on automatic identification of aggression for Russian and American imageboards. We have chosen 2ch.hk and 4chan.org as the most prominent and popular imageboards for their respective countries [C].

Aggression was detected by our algorithm based on the neural network library word2vec (Mikolov, Chen, Corrado, and Dean, 2013) and its Gensim (Řehůřek & Sojka, 2010) implementation for the Python programming language. Word2vec is an unsupervised algorithm that allows finding semantic relations and distances between words without any annotation or other data preprocessing. Nowadays this method is considered to be the best for determining semantic relations between words (Arefyev, Lesota, and Lukanin, n.d.). Although, some researchers argued that their systems performed better. For example, J. Pennington and R. Socher offered an algorithm called GloVe (Global Vectors for Word Representations) and proved that it outperforms word2vec (Pennington, Socher,