Abusive behavior in social media

Tarek Saier
tareksaier@gmail.com

1. INTRODUCTION

Usage of online platforms of all shapes and sizes nowadays is a common part of may people's everyday life. Just like human interaction offline, user interaction on Facebook, Twitter, online forums etc. is not always positive. For all the good like helpful contributions to Wikipedia and engaged discussion on reddit, there also is abusive behaviour taking place.

While the seriousness of the effects such behaviour can have on victims may have been downplayed in earlier days of the web, it is clearly as a serious problem. Furthermore, with the media reporting on large social networks failing to control abusive behaviour and thus influencing their public image, it is in the financial interest of companies running such networks to detect and remove or, if possible, even prevent such behavior.

This report will give an introduction into the topic of *Abusive behavior in social media* — or more precise: the decection of such behavior — and is structured as follows. Section 2 will give a wider view on the topic, provide necessary background information and shortly describe approaches for trackling the problem at hand. In section 3 the focus will be put on machine learning as one possible approach. While giving a short overview of the steps of a machine learning procedure in general, noteworthy particularities with regards to abusive behavior in social media will be explained. Section 4 will introduce two concrete approaches — efforts for detecting abusive comments on Yahoo! on the one hand and agressive Twitter accounts on the other. This will be followed by a comparison of the two. Lastly, section 5 will conclude the report.

2. BACKGROUND

In the physical world, abusive behavior can take many forms. Acts of bullying, for example, can be categorized into four types: physical, verbal, relational and damage to property[2]. In social media, physical abuse is not possible and damage to property at least rather unlikely and certainly not commonplace. While relational bullying is a possiblity, this report will focus on verbal types of abusive behaviour (that is, not only bullying but in general) from hereon.

2.1 Problem formulation

- What's the general problem
- What are the challenges (language changing, sarcasm, etc.)

2.2 Approaches for solving the problem

- Most basic: blacklisting of words
- More sophisticated: machine learning, deep learning, etc.

3. MACHINE LEARNING

- *short* description/recap of ML approach / noteworthy particularities with regards to topic at hand

3.1 Data collection

- No de facto testing set for abusive language[4]

3.2 Feature extraction

-

3.3 Learning

-

3.4 Evaluation

-

4. TWO CONCRETE APPROACHES

-

4.1 Abusive Yahoo! comments

- Description and discussion of [4]
  - NLP features (e.g. [3])
  - "Vowpal Wabbit's regression model"
  -

4.2 Aggressive Twitter accounts

- Description and discussion of [1]
  - WEKA, Random Forest
  -

4.3 Comparison

- How do [4] and [1] compare
  - Classifying accounts (more features) vs. just comments
  - Hate speech, derogatory language, profanity vs. bullying, aggression
  - Ground truth: trained staff vs. crowd sourcing
  -
- To what extend are they comparable

5. CONCLUSION

-

Seminar "Web Science", SS 2017

Albert-Ludwigs Universität Freiburg
Technische Fakultät, Institut für Informatik
Lehrstuhl für Datenbanken & Informationssysteme

6. REFERENCES

[1] D. Chatzakou, N. Kourtellis, J. Blackburn, E. D. Cristofaro, G. Stringhini, and A. Vakali. Mean birds: Detecting aggression and bullying on twitter. *CoRR*, abs/1702.06877, 2017.

[2] R. Gladden, A. Vivolo-Kantor, M. Hamburger, and C. Lumpkin. Bullying surveillance among youths: Uniform definitions for public health and recommended data elements, version 1.0, 2014.

[3] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. *CoRR*, abs/1405.4053, 2014.

[4] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 145–153, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.