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 behavior taking place.

While the seriousness of the effects such behavior 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 behavior 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[3]. 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 behavior in social media from hereon. Put simply, the remainder of this report is concerned with detecting absusive or malicious intent in text based communication.

Seminar "Web Science", SS 2017

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

2.1 Problem formulation

On a high level of abstraction, the task at hand is detecting and stopping abusive behavior in a social media setting. Within the scope of this report *"behavior"* boils down to communication in text form. This communication may be associated with user accounts. User accounts in turn also may have different qualities of interest.

Looking at it from the perspective of an entity operating a social platform, the problem can be formulated as: given all the information about the actors on our platform and the communication they engage in, how can abusive communication be detected? Depending on what constitues an actor on a platform, different possibilities for approaching that goal exist. For example, in the case of platform that wants its user accounts to be as representative of the real person controlling the account as possible (e.g. Facebook), it might be viable to detect abusive *accounts* in order to stop abusive communication. On the other hand, for a platform where the notion of an account does not hold much informational value, it might be more feasable to try and detect abusive communication from its contents only. Examples for the latter setting might be online comment sections that allow anonymous posting, or platforms like 2channel[1] or 4chan[2] that just give each participant an ID, so messages of the same origin can be identified as such, while connecting from a different IP address results in a different ID and the traceability of a common origin of messages is lost.

2.2 Challenges

As just described, the amount of information available can pose a challenge for detecting abusive behavior and render certain approaches nonviable. Aside from the mentioned feasablility to model user accounts, brevity of communication (e.g. on Twitter due to its 140 character limit) can also pose a hurdle.

Another set of challanges is given by the fact that the communication to be examined happens in natural language from the hands of humans. This means for example that:

- Offensive language may intentionally be obfuscated (e.g. *ni9 9er*) rendering simple keyword matching ineffective.
- Some language might be acceptable within one group of people but offensive within another.
- The offensive nature of an utterance might only come to light when considering a larger context (e.g. multiple sentences) while its parts taken out of context are harmless.

[1] `http://2ch.net/`
[2] `https://www.4chan.org/`

- Sarcasm might falsely be detected as abusive language, while constant sarcasm towards a user could also be a form of bullying.
- Language changing over time might require detection methods to be adapted over and over again.

Lastly, advancing the field of abusive behavior detection can be challenging because a lot of work may target different types of abusive behavior (e.g. bullying, aggression, hate speech, derogatory language, profanity) and therefore be incompatible.

2.3 Approaches for solving the problem

The following will give a very brief overview of some approaches to detect abusive language, while the remainder of the report will focus on machine learning as one such approach.

2.3.1 Most basic

An overly simplistic approach — nevertheless often seen applied in online forums — is to maintain a list of words that are deemed offensive, and filtering out those words or messages containing at least one blacklisted word. As described before this is easily circumvented by users that intentionally misspell or obfuscate offensive terms. Another problem with this approach is that a lot of harmless words containing offensive terms (e.g. to snigger) might falsely be flagged.

2.3.2 More sophisticated

More sophisticated approaches don't just take into account the words as isolated merely syntactical bits of information. They rather bring them into context and try to evaluate them semantically. A first step in this direction is, for example, considering the TFIDF scores of words. Furthermore, to allow for variations and obfuscated versions of words, not the words themselves but n-grams can be used[6]. Recent techniques most often are machine learning[5][2] or deep learning[1] approaches that take into account many different language-based, network-based and, if appilcable, user-based features.

3. MACHINE LEARNING

To develop a supervised machine learning approach for the task of abusive behavior detection one needs to:

- have or produce a data set with labels (e.g. a large set of messages labeled *abusive* or *not abusive*)
- decide on which features to extract from the data
- decide on a learning algorithm
- evaluate the system

The following sections will briefly describe each step and, if appilcable, highlight noteworthy particularities with regards to abusive behavior in social media.

3.1 Data collection

For the algorithm to work on, a large set of labeled data is necessary. This can be freely available and established testing sets (like the MovieLens data set[3] for the movie domain) or newly obtained data labeled via crowdsourcing or trained personnel. The data will be used to train and also evaluate the system. In the case of abusive behavior detection in social media, data is still problematic to a certain extend, since there is no de facto standard testing set for abusive language[5].

3.2 Feature extraction

The extraction of features from the data is essential in that it dictates what the learning algorithm gets as its input. This step might include *feature processing* steps — for example the generation of n-grams from whole words.

Common features in the case of abusive behavior detection in social media are text- or language-based (e.g. n-grams, results of seniment analysis and word embeddings), network-based (e.g. cliques and centrality scores) and user-based (e.g. account creation and activity times).

3.3 Learning

Once a labeled data set and its features are available, a learning algorithm can be trained with it to afterwards itself attempt to put correct labels on data.

3.4 Evaluation

The last step is to use the trained model for predicting labels and checking these predictions against the true labels. This ties in with the data collection step, which means for abusive behavior detection in social media the *creation* of a set with "true" labels from scratch is most propably necessary which makes the comparison of evaluation results of different approaches difficult.

4. TWO CONCRETE APPROACHES

The basis of this report are two papers, *Abusive language detection in online user content* by Nobata et al.[5] and *Mean birds: Detecting aggression and bullying on twitter* by Chatzakou et al.[2]. Sections 4.1 and 4.2 will describe and discuss them respectively and section 4.3 will compare the two.

4.1 Abusive Yahoo! comments

4.1.1 Task

In [5] Nobata et al. set out to detect comments on Yahoo! News[4] and Yahoo! Finance[5] that contain hate speech, derogatory language or profanity. For their definitions and examples of the three categories see figure 1.

4.1.2 Method

Data. The data sets are based on comments from Yahoo! News and Yahoo! Finance. In total 4 different sets are used. The first of these, the *primary data set* is comprised of a random 10% of the comments from the time between October 2012 and January 2014, plus all comments that have been reported as abusive by users during that time. In total there are a little over 2.1 million comments of which about 13% are labeled as abusive. For training and testing the data is split into parts of 80% and 20% respectively. Second, the *temporal data set* is sampled between April 2014 and April 2015. It consists of about 1.2 million comments of which a little over 7%[6] are labeled abusive. The *WWW2015 data set* is taken from a prior work were the comments are only taken from Yahoo! Finance during an unspecified 6 month

[3] `https://grouplens.org/datasets/movielens/`

[4] `https://www.yahoo.com/news/`

[5] `https://finance.yahoo.com/`

[6] According to the authors the significantly lower number compared to the *primary data set* is explained by the fact that their models were already in use on the platforms during that time.

<table>
<tr><td colspan="2">Given the passage below, first rate the passage on whether the passage is clean or abusive. And if abusive,
• whether it contains hate speech
• whether it contains derogatory languge
• whether it contains profanity</td></tr>
<tr><td><b>Hate Speech</b>:</td><td>language which attacks or demeans a group based on race, ethnic origin, religion, disability, gender, age, disability, or sexual orientation/gender identity. Some examples:
• The JEWZ are Market Manipulators Hate them freaking Jews !! To bad Hitler did not finish them off !!!
• I hate the mall because there are too many unemployed apefricans and mudsharks.
• 300 missiles are cool! Love to see um launched into TelAviv! Kill all the gays there!
• EGYPTIANS ARE DESCENDED FROM APES AND PIGS. NUKE EGYPT NOW Ah me tu se ku sh</td></tr>
<tr><td><b>Derogatory</b>:</td><td>language which attacks an individual or a group, but which is not hate speech. Some examples:
• All you perverts (other than me) who posted today, needs to leave the O Board. Dfasdfdasfadfs
• yikes...another republiCUNT weighs in....</td></tr>
<tr><td><b>Profanity</b>:</td><td>language which contains sexual remarks or profanity. Some examples:
• T.Boone Pickens needs to take a minimum wage job in FL for a week. I guarantee he shuts the f up after that.
• Somebody told me that Little Debbie likes to take it up the A.$.$.
• So if the pre market is any indication Kind of like the bloody red tampons that you to suck on all day??</td></tr>
</table>

Figure 1: Annotation instructions given to Yahoo! employees for labeling comments in [5].

period. There are almost 1 million comments, about 6% of which are labeled abusive. Lastly, the *evaluation data set*, a data set where extra effort was put into a valid labeling, is sampled and selected between March and April of 2015 and consists of 1000 clean and 1000 abusive comments.

Labelling. Trained employes and crowd sourcing experiments and eval data set extra effort.

- NLP features (e.g. [4])
- "Vowpal Wabbit's regression model"

4.1.3 Results

4.1.4 Discussion

- Mentions outperforming "state of the art deep learning approach" but never makes any referecnce to one

4.2 Aggressive Twitter accounts

In [2]

4.2.1 Task

4.2.2 Method

- WEKA, Random Forest

4.2.3 Results

4.2.4 Discussion

4.3 Comparison

- How do [5] and [2] 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

-

6. REFERENCES

[1] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 759–760, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.

[2] 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.

[3] 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.

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

[5] 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.

[6] D. Yin, B. D. Davison, Z. Xue, L. Hong, A. Kontostathis, and L. Edwards. Detection of harassment on web 2.0. In *Proceedings of the Content Analysis in the WEB*, 2009.