

Heidelberg University
Institute of Computer Science
Database Systems Research Group

Project Proposal for the lecture Text Analytics
Hate Speech Detection

Team Member: Christoper Klammt, 3588474, Applied Computer Science
iv249@stud.uni-heidelberg.de

Team Member: Felix Hausberger, 3661293, Applied Computer Science
eb260@stud.uni-heidelberg.de

Team Member: Nils Krehl, 3664130, Applied Computer Science
pu268@stud.uni-heidelberg.de

1 Motivation

Legal implementations on handling hate speech is different from one country to another. While hate speech is not prohibited in the United States due to the *freedom of speech*, other countries especially in the European Union can sue hate speech actors for either offending the public order or human dignity. While being able to prosecute actors in public without much effort, the internet and especially social media platforms provide an easy and anonymous way to practice hate speech without legal consequence enforcements. Several steps were taken to tackle hate speech online, one of them being the *code of conduct* on countering illegal hate speech online, an initiative of the European Commission in close collaboration with major IT companies like *Facebook*, *Microsoft*, *Twitter* and *YouTube* [1]. While respecting the freedom of speech, these companies commit to delete hate speech contributions within 24 hours of the initial deletion request. To further automatize the process of detecting hate speech contributions, several text analytics approaches have been evaluated in the recent past. Many of them are using methods of *natural language processing* and *deep learning* for hate speech detection and rely on meaningful features being learned automatically by neural networks instead of using hand-crafted features. In this work, the boundaries of conventional text analytics approaches for hate speech detection including manual feature selection and subsequent text classification of hate speech and non hate speech documents for the Twitter API should get evaluated. The result of this work should show which features work best for which classifier and which problems can be addressed with conventional text analytics methods and which not.

To add: -papers -datasets

2 Research Topic Summary

Current research in the area of hate speech detection is often built upon neuronal network techniques. Exemplary therefore are the works of Roy et al. [4], Setyadi, Nasrun and Setianingsih [5] and Kapil, Ekbal and Das [2], to name just a few examples. Among other metrics by considering the accuracy one can evaluate the success of a chosen approach. Roy et al. [4] have reached accuracies up to 95%.

Nevertheless even classical Machine Learning techniques such as Support Vector Machines (SVM) combined with text analytics methods are a promising approach. Watanabe, Bouazizi and Ohtsuki [6] used a decision tree and reached an accuracy of 87.4% for detecting whether a tweet is hateful or not.

Furthermore the results of classical Machine Learning techniques can be used as reference value to evaluate the neuronal network approach. This was done by Roy et al. [4].

There are three aspects to consider, derived from the domain of automatic hate speech detection:

1. Raw dataset
2. Feature extraction
3. Machine Learning technique

The following sections introduce these three aspects.

2.1 Raw dataset

To obtain a dataset one can use an existing already labelled dataset (e.g. done by Watanabe, Bouazizi and Ohtsuki [6]), or one could label the data by hand (e.g. made by Oriola and Kotzé [3]). Depending on the dataset the classes in which the data is classified can differ. Prominent classifications for datasets in the domain of hate speech are the binary classification (no hate speech, hate speech) and the ternary classification (clean, offensive, hate speech). TODO: Beispieldatensätze

Another aspect is the distribution between the different classes. Is the distribution equally among the different classes or not? In case of imbalanced distributions the machine learning approach can be less performant and accurate [3].

2.2 Feature extraction

The main difference between the two introduced methodologies (neuronal network, classical Machine learning techniques) is the process of feature extraction. In neuronal network approaches the used features are learned automatically, whereas classical Machine Learning techniques require a manual feature extraction process. This is highly based on text analytics.

The following list provides an overview of possible features:

- Sentiment-based features allow us to extract the polarity of the tweet
- Semantic features allow us to find any emphasized expression -> TF-IDF, Part-of-speech
- Unigram features allow us to detect any explicit form of hate speech -> N-grams, Bag-of-words

- Pattern features allow the identification of any longer or implicit forms of hate speech

TODO: Auflistung welche Features in welchen Papern schon betrachtet wurden.

2.3 Machine Learning technique

Welche ML methods werden benutzt

3 Project Description