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 automize 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

There are three aspects to consider, derived from the reseach on automatic hate speech detection:

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

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 [2]), 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).

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

There are two classes of machine learning approaches for detecting hate speech (neuronal network approaches, classical machine learning approaches). They are described in detail in the next chapter. One main difference between the two introduced approaches 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 and their technical methods from the area of text analytics. The list is clustered into the four categories based on Watanabe, Bouazizi and Ohtsuki [2]. The possible text analytics approaches derive from the literature review of Fortuna and Nunes [4].

- **Sentiment-based features**: Is the tweet rather positive or negative?
  Text analytics approaches: Dictionaries, Rule Based Approaches, Objectivity-Subjectivity of the Language, Declarations of Superiority of the Ingroup [4]

- **Semantic features**: Which parts of the tweet are emphasized?
  Text analytics approaches: TF-IDF, Part-of-speech, Profanity Windows, Lexical Syntactic Feature-based, Topic Classification, Template Based Strategy, Word Sense Disambiguation Techniques, Othering Language [4]

- **Unigram features**: Are there any specific words marking hate speech?
  Text analytics approaches: N-grams, Bag-of-words [4]

- **Pattern features**: Are there any specific patterns marking hate speech? Text analytics approaches: Part-of-speech, Dictionaries, Typed Dependencies,Word Embeddings [4]

2.3 Machine learning approach

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. [5], Setyadi, Nasrun and Setianingsih [6] and Kapil, Ekbal and Das [7], to name just a few examples. Among other metrics by considering the accuracy one can evaluate the success of a choosen approach. Roy et al. [5] have reached accuracies up to 95% for detecting whether a tweet is hateful or not.

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 [2] used a decision tree and reached an accuracy of 87.4%.

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. [5].

3 Project Description

This chapters aim is to introduce the chosen project approach. First of all the main project goals are presented. Based on the selected raw dataset the planned text analytics tasks and the corresponding pipeline are introduced. Finally the used machine learning approaches and their evaluation are presented.

3.1 Project goals

3.2 Raw dataset

3.3 Text Analytics Tasks and Pipeline (Feature extraction)

3.4 Machine learning approach and Evaluation

References

[1] European Commission, *The code of conduct on countering illegal hate speech online*, European Commission Homepage, 2020. [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_1135 (visited on 11/19/2020) (cit. on p. 1).

[2] Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki, *Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection*, 2018. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8292838 (visited on 11/23/2020) (cit. on pp. 2, 3).

[3] Oluwafemi Oriola and Eduan Kotzé, *Evaluating machine learning techniques for detecting offensive and hate speech in south african tweets*, 2020. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8963960 (visited on 11/23/2020) (cit. on p. 2).

[4] Paula Fortuna and Sérgio Nunes, *A survey on automatic detection of hate speech in text*, 2018. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3232676 (visited on 11/23/2020) (cit. on pp. 2, 3).

[5] Pradeep Kumar Roy, Asis Kumar Tripathy, Tapan Kumar Das, and Xiao-Zhi Gao, *A framework for hate speech detection using deep convolutional neural network*, 2020. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9253658 (visited on 11/23/2020) (cit. on p. 3).

[6] Nabiila Adani Setyadi, Muhammad Nasrun, and Casi Setianingsih, *Text analysis for hate speech detection using backpropagation neural network*, 2018. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8712109 (visited on 11/23/2020) (cit. on p. 3).

[7] Prashant Kapil, Asif Ekbal, and Dipankar Das, *Investigating deep learning approaches for hate speech detection in social media*, 2020. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/2005/2005.14690.pdf (visited on 11/23/2020) (cit. on p. 3).