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 machine learning approaches for hate speech detection should get evaluated including manual feature extraction and subsequent text classification of hate speech and non-hate speech documents. The data sets used in this work originate from *Twitter* posts [2] and contributions to the *White Supremacy Forum* [3]. The result of this work should show which features work best for which classifier and which problems can be addressed with conventional machine learning methods and which not as opposed to deep learning approaches. In the first part results of the research phase will be presented before moving on to the conrete project description.

2 Research Topic Summary

The main contribution of this work should be to evaluate the boundaries of conventional machine learning approaches for hate speech detection including manual feature extraction and subsequent text classification of documents into hate speech and non hate speech as opposed to modern deep learning approaches. The results should show which features work best for which classifier and which problems can be addressed with conventional machine learning methods and which not. A comparison to modern deep learning approaches should be drawn using the metrics of simple accuracy as well as

precision and recall, respectively the F1-score.
In order to achieve the goals of this work three milestones were identified:

1. Data preparation and representation
2. Feature extraction
3. Conventional machine learning approaches

The following sections introduce these three aspects.

2.1 Data preparation and representation

There are two data sets used for the project. The first one [2] uses data from the *Twitter API*[1]. It consists of a sample of around 25k tweets that were identified as hate speech based on a previously composed hate speech lexicon without regarding context information. Subsequently, each document in the corpus got labeled with one of the three categories *hate speech*, *offensive language* or *neutral*. Therefore the data set follows a classical ternary classification style. The workers were adviced to follow predefined definitions of each category and to take context information into consideration. Each tweet was coded by three or more workers. The majority of tweets were classified as offensive language (76% at 2/3, 53% at 3/3), only 5% were coded as hate speech. The data is provided offline as a CSV or pickle file.
The second data set uses data from the *White Supremacy Forum* [3][2]. One document represents a sentence that is according to binary classification either labeled as hate or no hate. In total, 1.119 sentences containing hate and 8.537 sentences containing no hate are provided. Once again the documents were labeled manually by human actors following previously specified guidelines, on request additional context information were provided. The documents are given offline as normal text files, annotations are stored in a CSV file.
Both data sets are stated to be balanced, multiple documents cannot be traced back to a single user. In case of imbalanced distributions classifications can be less performant and accurate [4].

2.2 Feature extraction

There are two approaches for detecting hate speech, either using statistical and probabilistic methods from a conventional machine learning background

[1] https://github.com/t-davidson/hate-speech-and-offensive-language
[2] https://github.com/Vicomtech/hate-speech-dataset

or using deep learning based approaches. One main difference between the two approaches is the process of feature extraction. In deep learning approaches the used features are learned automatically, whereas classical text analytics techniques require a manual feature extraction process.
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 [5]. The possible text analytics approaches derive from the literature review of Fortuna and Nunes [6].

- **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 [6]

- **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 [6]

- **Unigram features**: Are there any specific words marking hate speech?
  Text analytics approaches: N-grams, Bag-of-words [6]

- **Pattern features**: Are there any specific patterns marking hate speech?
  Text analytics approaches: Part-of-speech, Dictionaries, Typed Dependencies,Word Embeddings [6]

2.3 Conventional machine learning approaches

Current research in the area of hate speech detection is often built upon deep learning techniques. Exemplary therefore are the works of Roy et al. [7], Setyadi, Nasrun and Setianingsih [8] and Kapil, Ekbal and Das [9], 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. [7] have reached accuracies of up to 95% for detecting whether a tweet is hateful or not.
Nevertheless even classical machine learning techniques such as Support Vector Machines combined with text analytics methods are a promising approach. Watanabe, Bouazizi and Ohtsuki [5] used a decision tree and reached an accuracy of 87.4%. In another paper Oriola and KotzÃ© [4] have shown

that optimized gradient boosting with word n-gram can achieve a true positive rate of 86.7%. Gaydhani, Doma, Kendre and Bhagwat [10] even evaluated various machine learning models and based on n-grams and their according TF-IDF values and achieved an accuracy of 95.6% which can definitely compete with modern deep learning approaches until a certain threshold. Other relevant work was documented on GitHub[3].

When being restricted to a binary classification model one can easily use the concepts that will be proven by this work and apply it to other binary classification tasks like spam detection.

3 Project Description

github

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

[3]https://github.com/fidsusj/HateSpeechDetection

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] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber, *Automated hate speech detection and the problem of offensive language*, 2020. [Online]. Available: https://arxiv.org/pdf/1703.04009.pdf (visited on 11/23/2020) (cit. on pp. 1, 2).

[3] Ona de Gibert, Naiara Perez, Aitor Garcia-Pablos, and Montse Cuadros, *Hate speech dataset from a white supremacy forum*, 2020. [Online]. Available: https://www.aclweb.org/anthology/W18-5102.pdf (visited on 11/23/2020) (cit. on pp. 1, 2).

[4] 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 pp. 2, 3).

[5] 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 p. 3).

[6] 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 p. 3).

[7] 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).

[8] 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).

[9] 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).

[10] Aditya Gaydhani, Vikrant Doma, Shrikant Kendre, and Laxmi Bhagwat, *Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach*, 2018. [Online]. Available: `https://arxiv.org/pdf/1809.08651.pdf` (visited on 11/23/2020) (cit. on p. 4).