Heidelberg University
Institute of Computer Science
Database Systems Research Group

Report for the lecture Text Analytics

Hate Speech Detection

`https://github.com/fidsusj/HateSpeechDetection`

Mentor: John Ziegler

Team Member:	Christopher 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

Abstract

Plagiarism statement

We certify that this report is our own work, based on our personal study and/or research and that we have acknowledged all material and sources used in its preparation, whether they be books, articles, reports, lecture notes, and any other kind of document, electronic or personal communication. We also certify that this report has not previously been submitted for assessment in any other unit, except where specific permission has been granted from all unit coordinators involved, or at any other time in this unit, and that we have not copied in part or whole or otherwise plagiarised the work of other students and/or persons.

Contents

1 Introduction

One current research area in the field of text analytics is hate speech detection. Many approaches from the recent past take use of neural network architectures to deal with such classification problems. One of the most common approaches is to use uni- and bidirectional long short-term memory (LSTM) networks, a recurrent neural network architecture that can process input of arbitrary length and remembers context information [6, 16, 15]. The paper [8] states, that even a simple gated recurrent unit (GRU) architecture can perform as good as more complex units. [1] repurposes the famous bidirectional encoder representations from transformers (BERT) language model to perform classification tasks for hate speech detection. Besides that, other approaches use convolutional neural networks (CNNs) to extract typical hate speech patterns [2, 14, 10] or even deep belief network algorithms [12]. Using neural network approaches means to automatically learn representative features for the classification task. On the other hand, the papers introduced in section 2 use a different approach by solving the classification task with manually extracted features. Nevertheless, none of the papers combines the different achievements of such recent research and compares it to a baseline neural network architecture, which is what this work is dedicated to.

After a definition of the term "hate speech" in section 3 different classifiers will be trained on a holistic, hand-crafted feature set based on recent publications in the field of hate speech detection. The task includes building and preprocessing a training corpus as well as introducing and explaining the different kinds of features. How well the different classifiers perform compared to a neural network approach as a baseline and several statictical insights into typical hate speech artifacts will be presented in section 4. The results of this work in section 5 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 neural network approaches. A summary over the achievements earned will be drawn in section 6.

2 Related Work

As the goal of this work is to solve a hate speech classification problem with manually extracted features, recent work was evaluated proposing different feature sets for this task.

[17] categorizes features into four different groups. Sentiment-based features give information about the polarity of a document, which is important as many hate speech documents stand out by being mostly negativ. Sentiment features count the occurences of punctuations, capitalized words, interjections, etc. A dictionary of typical hate speech words can be obtained by extracting most common unigrams from a given corpus uilding thr unigram features. Pattern features represent the last feature group containing common syntactic patterns based on PoS tags. The apporach presented in this paper achieved an accuracy of 87.4% using combined features from these four groups. The classifiers used were SVM, Random Forest and J48graft.

[13] concludes character n-grams, word n-grams, negative sentiment-based scores and syntactic-based features as a decent feature set to train classifiers on. Watching at the results, an optimized support vector machine with character n-grams performed best with 0.894 TPR, while optimized gradient boosting performed best with word n-grams, giving a 0.867 TPR.

[7] divides features into generic text mining features and specific hate speech detection features. Typical generic text mining features are dictionaries of insults typical for heate speech, swear words, profane words verbal abuse, etc., n-grams, lexical syntactic based template features, that capture grammatical dependencies within a sentence, topic classifications with latent dirichlet allocation, or sentiment polarity scores. On the other hand specific hate speech detection features do not rely on common abstract concepts known in the field of text analytics, but come with purpose built frameworks to detect these featues. Using the Stanford lexical parser along with a context-free lexical parsing model one can identify othering language which used a lot in hate speech. Other examples of specific hate speech features are the objectivity-subjectivity relations of the language as hate speech is more related to subjective communication, focus on particular stereotypes, intersectionism of oppression or declarations of superiority of the ingroup.

Other related work like [9], [11] and [4] once more stress the importance of word and character n-grams for hate speech detection tasks. [4] even uses count indicators for hashtags, mentions, retweets and URLs and is esspecially important as a part of the dataset used in this work originated from the project behind this paper. The best performing model achieved 0.91% overall precision, 0.9% recall and a 0.9 F1-score, but the model is biased towards classifying tweets as less hateful or offensive than the human supervisors.

3 Approach

3.1 Definition of Hate Speech

3.2 Dataset

3.3 Preprocessing

3.4 Definition of feature groups

3.5 Classifiers

3.6 Evaluation

4 Experimental setup and results

4.1 Data

There are two data sets used for the project. The first one uses data from the *Twitter API* [4][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 instructed to follow predefined definitions of each category and to take context information into consideration. Each tweet was assessed and labeled 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* [5][2]. One document represents a sentence that is either labeled as hate or not hate. In total, 1.119 sentences containing hate and 8.537 sentences being non-hate are provided. Once again the documents were labeled manually by human actors following previously specified guidelines, on request additional context information was provided. The documents are given offline as normal text files with annotations stored in a separate CSV file.

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

[2] https://github.com/Vicomtech/hate-speech-dataset

4.1.1 Data preparation

To prepare a central dataset, both single datasets had to be transformed into a common format. For the central dataset only the class and the text content of each tweet respectively each forum contribution was considered.

The first dataset “Automated Hate Speech Detection and the Problem of Offensive Language” was entirely given as a .csv file and contains 25.297 tweets, that were either labeled as hate speech, offensive language or neither of both. To determine the right label three independent evaluators classified each tweet, the final label got assigned by the majority vote. As for the first approach, one is only interested in hate speech and neutral tweet classification, all offensive language documents in the dataset were dropped. Some tweets were retweets that were commented additionally by a user. As it could not be distinguished whether the original tweet or the retweet contains hate speech, these documents were filtered out as well. An example is shown below:

> """@jaimescudi_: ""@Tonybthrz_: ""@jaimescudi_: I swear if oomf try talking to me tomorrow.."" @"" @BarackObama"" pussy"

The original tweets can be found in between the ""..."". Same goes for tweets that cite other users without using the retweet option.

The second dataset **OnadeGibert2020** was not entirely given as a .csv file. Only the document annotations were given in a .csv file, all forum contributions were stored in separate .txt files. Only documents which could not be assigned to a single class (label "idk/skip") or referred to other documents (label "relation") were dropped.

The resulting common dataset was stored in a .csv file. It contains 2.491 hate speech documents and 13.336 non hate speech documents. The dropped offensive language documents make up 17.505 instances. In case the classification results are too poor, additional 2.818 offensive language documents can be added that were labeled as hate speech by one evaluator.

4.1.2 Corpus building

The common dataset is loaded from the .csv file into a pandas dataframe. After doing basic preprocessing like removing emojis and other irrelevant characters, spacy is used to build a tokenized corpus. The language model that spacy brings decides about stop word, punctuation and white space removal. No hard coded logic or stop word lists are used in this process. This keeps URLs or other tokens including punctuation as one token. Furthermore no stemming was applied to the tokens, instead lemmatization was used as one can in this case later on use pre-trained word embeddings from i.e. Word2Vec. Furthermore tokenization works better using the lemmas instead of word stems (e.g. We'll becomes ["we","will"] and not ["we", "'ll"].

4.1.3 Data analysis

As already mentioned, the acquired dataset is an imbalanced one, which can lead to a decrease in performance and accuracy with machine learning classification. As a comparison the paper [13] also recognizes the class imbalance and tries to reduce it by applying a synthetic minority oversampling technique called SMOTE [3]. In general there are a few possibilities to tackle the challenge of unbalanced classes:

- changing the performance metric (e.g. F1-score instead of accuracy)
- undersampling, i.e. deleting instances from the over-represented class
- oversampling, i.e. adding copies of instances from the under-represented class
- generating synthetic samples (e.g by using SMOTE)

In our experiments we chose to try the different approaches and applied a simple undersampling, as well as an oversampling using SMOTE. Additionally, we used the imbalanced dataset to train a classifier to compare how this affects the performance.

In further analysis of the data, we had a look at the length of hate speech posts versus non-hate speech posts. This can be seen in Figure 1.

Here one can see, that the hate speech posts contain more words (tokens before cleaning) than non-hate speech posts. In average a hate speech post contains 18.18 words, whereas a non-hate speech post only contains 15.85 words. Unlike expected, the hate speech posts are longer than the non-hate speech posts.

A more interesting look at the data are the most commonly used words per class. As can be seen in the word clouds in Figure 2 there are some obvious differences, such that the hate speech posts use words like “bitch”, “faggot” or “nigga”. But interestingly enough, the non-hate speech posts also often consist of the words “trash” or “white”.

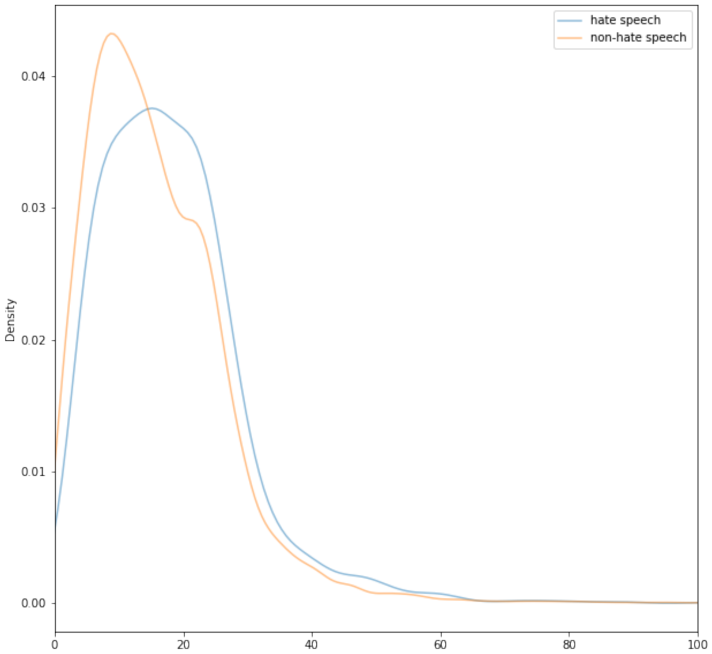


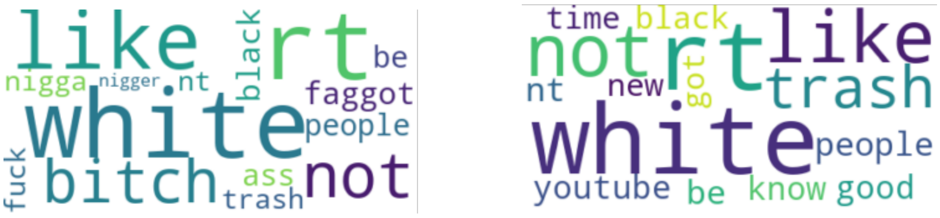
Figure 1: Density distribution of the length of a post (tweet or forum post)

For a better insight with what data we are dealing, a few examples are shown in the following.
Examples for non-hate speech (neutral sentences):

- "billy that guy would nt leave me alone so i gave him the trudeau salute"
- "this is after a famous incident of former prime minister pierre trudeau who gave the finger to a group of protesters who were yelling antifrench sayings at him"
- "askdems arent you embarrassed that charlie rangel remains in your caucus"

Examples for hate speech:

- "california is full of white trash"



(a) hate speech

(b) non-hate speech

Figure 2: Word clouds

- "and yes they will steal anything from whites because they think whites owe them something so it s ok to steal"
- "why white people used to say that sex was a sin used to be a mystery to me until i saw the children of browns and mixed race children popping up all around me"

One can clearly see the hate expressed in the hate speech examples and see their discriminating nature.

5 Analysis

6 Conclusion

References

[1] Hind Saleh Alatawi, Areej Maatog Alhothali, and Kawthar Mustafa Moria. “Detecting White Supremacist Hate Speech using Domain Specific Word Embedding with Deep Learning and BERT”. In: *arXiv* (2020). URL: https://arxiv.org/abs/2010.00357.

[2] Pinkesh Badjatiya et al. “Deep Learning for Hate Speech Detection in Tweets”. In: *Proceedings of the 26th International Conference on World Wide Web Companion* (2017). DOI: 10.1145/3041021.3054223.

[3] Nitesh V. Chawla et al. “SMOTE: Synthetic Minority Over-sampling Technique”. In: *arXiv* (2011). URL: https://arxiv.org/abs/1106.1813.

[4] Thomas Davidson et al. “Automated Hate Speech Detection and the Problem of Offensive Language”. In: *arXiv* (2017). URL: https://arxiv.org/abs/1703.04009.

[5] Ona De Gibert et al. “Hate Speech Dataset from a White Supremacy Forum”. In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)* (2018). DOI: 10.18653/v1/W18-5102.

[6] Wyatt Dorris et al. “Towards Automatic Detection and Explanation of Hate Speech and Offensive Language”. In: *Proceedings of the Sixth International Workshop on Security and Privacy Analytics* (2020). DOI: 10.1145/3375708.3380312.

[7] Paula Fortuna and SÃ©rgio Nunes. “A Survey on Automatic Detection of Hate Speech in Text”. In: *ACM Computing Surveys* (2018). DOI: 10.1145/3232676.

[8] Antigoni Maria Founta et al. “A Unified Deep Learning Architecture for Abuse Detection”. In: *Proceedings of the 10th ACM Conference on Web Science* (2019). DOI: 10.1145/3292522.3326028.

[9] Aditya Gaydhani et al. “Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach”. In: *arXiv* (2018). URL: https://arxiv.org/abs/1809.08651.

[10] Prashant Kapil, Asif Ekbal, and Dipankar Das. “Investigating Deep Learning Approaches for Hate Speech Detection in Social Media”. In: *arXiv* (2020). URL: https://arxiv.org/abs/2005.14690.

[11] Shervin Malmasi and Marcos Zampieri. "Detecting Hate Speech in Social Media". In: *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)* (2017). DOI: 10.26615/978-954-452-049-6_062.

[12] Iqbal Zulfikar Muhammad, Muhammad Nasrun, and Casi Setianingsih. "Hate Speech Detection using Global Vector and Deep Belief Network Algorithm". In: *1st International Conference on Big Data Analytics and Practices (IBDAP)* (2020). DOI: 10.1109/IBDAP50342.2020.9245467.

[13] Oluwafemi Oriola and Eduan KotzÃ©. "Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets". In: *IEEE Access* 8 (2020). DOI: 10.1109/ACCESS.2020.2968173.

[14] Pradeep Kumar Roy et al. "A Framework for Hate Speech Detection Using Deep Convolutional Neural Network". In: *IEEE Access* 8 (2020). DOI: 10.1109/ACCESS.2020.3037073.

[15] Arum Sucia Saksesi, Muhammad Nasrun, and Casi Setianingsih. "Analysis Text of Hate Speech Detection Using Recurrent Neural Network". In: *International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC)* (2018). DOI: 10.1109/ICCEREC.2018.8712104.

[16] Syahrul Syafaat Syam, Budhi Irawan, and Casi Setianingsih. "Hate Speech Detection on Twitter Using Long Short-Term Memory (LSTM) Method". In: *4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)* (2019). DOI: 10.1109/ICITISEE48480.2019.9003992.

[17] 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". In: *IEEE Access* 6 (2018). DOI: 10.1109/ACCESS.2018.2806394.