Heidelberg University
Institute of Computer Science
Database Systems Research Group

Project Proposal for the lecture Text Analytics

Hate Speech Detection

Team Member: Felix Hausberger, 3661293, Applied Computer Science
eb260@stud.uni-heidelberg.de
Team Member: Christoper Klammt, 3588474, Applied Computer Science
iv249@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 where 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

- show selection of DL papers

3 Project Description

References

[1] European Commission. The code of conduct on countering illegal hate speech online, 2020.