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 [1, 2, 3]. The paper [4] states, that even a simple gated recurrent unit (GRU) architecture can perform as good as more complex units. [5] 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 [6, 7, 8] or even deep belief network algorithms [9]. 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

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

5 Analysis

6 Conclusion

References

[1] Dorris, W., Hu, R., Vishwamitra, R., Luo, F., Costello, M.: Towards Automatic Detection and Explanation of Hate Speech and Offensive Language. In: Proceedings of the Sixth International Workshop on Security and Privacy Analytics (2020). https://doi.org/10.1145/3375708.3380312

[2] Syam, S., Irawan, B., Setianingsih, C.: 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). https://doi.org/10.1109/ICITISEE48480.2019.9003992

[3] Saksesi, A., Nasrun, M., Setianingsih, C.: Analysis Text of Hate Speech Detection Using Recurrent Neural Network. In: International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC) (2018). https://doi.org/10.1109/ICCEREC.2018.8712104

[4] Founta, A., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., Leontiadis, I.: A Unified Deep Learning Architecture for Abuse Detection. In: Proceedings of the 10th ACM Conference on Web Science (2019). https://doi.org/10.1145/3292522.3326028

[5] Saleh, H., Alhothali, A., Moria, K.: Detecting White Supremacist Hate Speech using Domain Specific Word Embedding with Deep Learning and BERT. In: arXiv (2020). https://arxiv.org/abs/2010.00357

[6] Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep Learning for Hate Speech Detection in Tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion (2017). https://doi.org/10.1145/3041021.3054223

[7] Roy, P., Tripathy, A., Das, T., Gao, X.: A Framework for Hate Speech Detection Using Deep Convolutional Neural Network. In: IEEE Access (Volume: 8) (2020). https://doi.org/10.1109/ACCESS.2020.3037073

[8] Kapil, P., Ekbal, A., Das, D.: Investigating Deep Learning Approaches for Hate Speech Detection in Social Media. In: arXiv (2020). https://arxiv.org/abs/2005.14690

[9] Muhammad, I., Nasrun, M., Setianingsih, C.: Hate Speech Detection using Global Vector and Deep Belief Network Algorithm. In: 1st International Conference on Big Data Analytics and Practices (IBDAP) (2020). https://doi.org/10.1109/IBDAP50342.2020.9245467