

Abusive behavior in social media

Tarek Saier
tareksaier@gmail.com

1. INTRODUCTION

- General introduction
- Why/to what extend is abusive behavior in social media a problem
- Structure of this report

2. BACKGROUND

- Wider overview before concentrating on ML afterwards

2.1 Problem formulation

- What's the general problem
- What are the challenges

2.2 Approaches for solving the problem

- Most basic: blacklisting of words
- More sophisticated: machine learning, deep learning, etc.

3. MACHINE LEARNING

- *short* description/recap of ML approach / noteworthy particularities with regards to topic at hand

3.1 Data collection

- No de facto testing set for abusive language[3]

3.2 Feature extraction

-

3.3 Learning

-

3.4 Evaluation

-

4. TWO CONCRETE APPROACHES

-

4.1 Abusive Yahoo! comments

- Description and discussion of [3]
- NLP[2]

4.2 Aggressive Twitter accounts

- Description and discussion of [1]

4.3 Comparison

- How do [3] and [1] compare
- To what extend are they comparable

5. CONCLUSION

-

6. REFERENCES

- [1] D. Chatzakou, N. Kourtellis, J. Blackburn, E. D. Cristofaro, G. Stringhini, and A. Vakali. Mean birds: Detecting aggression and bullying on twitter. *CoRR*, abs/1702.06877, 2017.
- [2] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. *CoRR*, abs/1405.4053, 2014.
- [3] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 145–153, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.