Implementation of a Unified Deep Learning Architecture for Abuse Detection

Mahdi Rahimi

February 13, 2019

1 Abstract / Introduction

The social media such as Twitter and online forums have been growing rapidly for the past years. This has provided a potential for propagation of abusive behavior such as hate speech, offensive, sexist and racist language, aggression, cyberbullying, harassment, and trolling. The fact that people who exploit from this opportunity can remain anonymous and their location would be unknown in the cyberspace, contributes even more to this problem and can even eventually lead to hate crimes beyond the reach of law enforcement.

Previously, custom-made approaches to address these abusive languages have been proposed. However, they have been tuned to detect only one specific type of abusive behavior. In this project, I would like to implement an approach by Founta et al [1] which is a more holistic approach and considers different aspects of these behaviors. The project is focused on Twitter and introduces a deep learning architecture which is two-fold. It utilizes a wide range of available meta data and combines it with a feature extraction method based on Recurrent Neural Networks which learns underlying characteristics of the tweets to recognize the abusive behavior. The project applies this architecture to four different types of abusive behavior: hate speech, sexism vs. racism, bullying, and sarcasm without the need to tuning the architecture for each task. For each of these four types, we use a separate and distinct dataset which address a certain abusive behavior. The results in the paper are higher than the state-of-art methods. I plan for my implementation to obtain a results that are as good as or better than them.

2 Describe your project in one sentence

In this project, I will implement an approach presented in [1] to detect several different types of abusive languages in online platforms.

3 Who is the audience for this project? How does it meet their needs? What happens if their needs remain unmet?

The problem of intensive abusive behavior has recently emerged. Unfortunately, it is a difficult problem to handle. Media platforms, such as Twitter and Facebook, host controversial contents, but they have not sufficiently addressed this problem. Currently, they are in a constant competition with abusive persons who constantly change their methods to bypass the detection algorithms. The method presented in this project is a universal and lightweight solution which can consider many existing meta data and detect different type of abusive language.

4 What is your approach and why do you think it will be successful?

The approach is to have a single model/architecture that incorporates domain specific metadata, as well as text content and is performant on a large number of abusive content as presented in [1]. The paper reports successful results.

5 In the best-case scenario, what would be the impact statement (conclusion statement) for this project?

The approach detects several types of abusive languages in a holistic way and with a performance that is better than the state-of-the-art.

6 List all major milestones for this project, and how you intend to spread them throughout the course.

1. Obtaining all the four datasets used in the paper
2. Preprocessing and data cleaning
3. Implementing the architecture
4. Training the model
5. Experimenting with four different datasets. I may not experiment with all of the four datasets if there is not enough time.

7 What obstacles do you anticipate, and how do you plan to address them?

7.1 Major obstacles

- I see no major obstacles.

7.2 Minor obstacles

- One of the four datasets is not available online and I have to contact the authors to obtain it.
- I need to find a proper computing resourses for the computationally intensive training.
- As mentioned earlier, I may not experiment with all of the four datasets if there is not enough time.

8 What additional resources do you need to complete this project?

- Learning the concept of ”attention” in NLP.
- Learning TensorFlow
- Finding computing resources

9 List 5 major publications that are most relevant to this project, and how they are related.

- As noted earlier, this project is the implementation of the works by Founta et al [1]
- Cyberbullying Dataset by Chatzakou et al [2]
- Offensive Dataset by Waseem et al [3]
- Hate Dataset by Davidson et al [4]
- Sarcasm Dataset by Rajadesingan et al [5]

10 When / How do you know if you have succeeded in this project?

I know I have succeeded in this project when I successfully do all the milestones and obtain good results.

11 References

[1] A.-M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis. A Unified Deep Learning Architecture for Abuse Detection. arXiv preprint arXiv:1802.00385, 2018.

[2] Chatzakou, D.; Kourtellis, N.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; and Vakali, A. 2017. Mean birds: Detecting aggression and bullying on twitter. In 9th ACM WebScience.

[3] Waseem, Z., and Hovy, D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In SRW@ HLT-NAACL, 88–93.

[4] Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. arXiv preprint arXiv:1703.04009.

[5] Rajadesingan, A.; Zafarani, R.; and Liu, H. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In 8th ACM WSDM, 97–106