Fighting Online Abuse and Bullying

Hasil Sharma[+], Paramjot Singh[+], Ajay Kedia[+], Ganesh Chandra[+], and Akshit Arora[+]

[+]University of Colorado Boulder, Department of Computer Science, Boulder CO, 80303, US
[+]these authors contributed equally to this work

ABSTRACT

Cyberbullying is defined as the deliberate and repeated harm inflicted through the use of cell phones/smartphones, computers/tablets, and other electronic devices (including Wi-Fi gaming devices). With applications like Twitter, Facebook, Medium etc. it has become easier for bullies to harass anyone from any part of the world. For many people, such online harassment is a nuisance — a sad reminder that the internet contains some of the worst parts of humanity along with the best. It creates unnecessary annoyance to those who are bullied and precludes them from enjoying cyber resources in best possible way. This project aims to build a machine learning system to understand the nuances and context of abusive language, determining incivility to help online content moderators keeping incidents such as abusive language, threats and harassment in check.

Why is this project worth studying?

In today's well connected world with platforms such as Twitter, Facebook, Medium etc. available at our disposal it is now easier than ever to voice your opinions and share your thoughts. A report from the Cyberbullying Research Center, released in 2016, indicates that 33.8% of students between 12 and 17 were victims of online harassment in their lifetime. Conversely, 11.5% of students between 12 and 17 indicated that they had engaged in cyberbullying in their lifetime. With these platforms cases of online harassment, cyber bullying etc. are also increasing. All these are a nuisance and precludes people from having a civil discourse.

Although the benefits of mobile information access are acknowledged through the empowering influence over its audience, a concern is noted with reference to largely uncensored forums offering mobile communication exchange to children. The proliferation of mobile technologies available, in conjunction with applications facilitating social networking, has steadily increased the attack surface minors are exposed to in an online environment. Most minors engaging in online activities do so through mobile technologies such as the cell phone. This device, as a consequence of its mobility, offers access to the internet that circumvents controls of supervision[1].

Why Machine Learning Techniques are appropriate?

Through machine learning, we can detect language patterns (such as grammatical structure, linguistics information etc.) used by bullies and their victims, and develop rules to automatically detect cyber bullying content. Most of the abusive/inappropriate sentences have a common underlying pattern in terms of particular words, grammatical structure, semantic inference etc. These patterns can be learned with appropriate machine learning techniques and can be used to gauge the civility of comment. We are planning to explore Neural Networks especially the class of Recurrent Neural Network (RNNs). RNN architectures have widespread real life applications such as Time Series Prediction, Music Composition, Speech Recognition etc. Given the capability to handle context and long-term dependencies, RNNs are widely used in Natural Language Processing domain. However, RNNs suffer from a major drawback: the context attenuates as the number of time-steps increase per input example. In order to address this limitation there are techniques such as Long Short Term Memory (LSTM), Gated Recurrent Units (GRU), Attention Mechanism etc., which we wish to learn more about.

Dataset(s)

Cyber harassment detection is an active area of interest in multiple research laboratories across the globe such as Kaspersky Labs, MIT Media Labs, etc. One of such labs, Conversation AI, has made their dataset public and we are planning on using same in our project. The data was originally collected from Wikipedia labeled talk where each comment is labeled on gender, ages, education, first-language, year, type of article (namespace), type of attack (quoting-attack, recipient-attack, third party-attack and other attack) and every comment is rated by at most 10 workers. We have created a kaggle notebook for the dataset.

We also have other datasets available like CU's CyberSafety Research Center, Formspring.me[2].

Our Approach

We are targeting to use vanilla RNN as our initial baseline and later incorporate model enhancement techniques to boost our model performance. We are planning using to use pre-trained word2vec models in order to efficiently encode the training examples. We plan to feed the learned features from Neural Network Architecture into a model to gauge the civility of the comment. There are multiple neural network and deep learning libraries such as Keras, Tensorflow etc. available out there which have these model or some part of it pre-implemented, and we plan on using these in our project. We can also use other techniques described here[3].

More specifically, unlike previous approaches(examine some of the papers) that considered a fixed window of a cyber-predator's questions within a dialogue, we will be exploiting the whole comment and model it as a signal, whose magnitude depends on the degree of bullying content. Using feature weighting and dimensionality reduction techniques, each signal will be straightforwardly parsed by a neural network that forecasts the level of insult within a comment given a window between two and three similar comments. By applying SVD on the time series data and taking into account the second dimension (since the first is usually modeling trivial dependencies between instances and attributes) we will be observing that its plot was very similar to the plot of the class attribute[4]. By applying a Dynamic Time Warping algorithm, the similarity of the aforementioned signals will be proved to exist, providing an immediate indicator for the severity of cyber bullying within a given dialogue.

References

1. Serra, S. M. & Venter, H. S. Mobile cyber-bullying: A proposal for a pre-emptive approach to risk mitigation by employing digital forensic readiness. In *2011 Information Security for South Africa - Proceedings of the ISSA 2011 Conference* (2011). DOI 10.1109/ISSA.2011.6027507.

2. Reynolds, K., Kontostathis, A. & Edwards, L. Using machine learning to detect cyberbullying. In *Proceedings - 10th International Conference on Machine Learning and Applications, ICMLA 2011*, vol. 2, 241–244 (2011). DOI 10.1109/ICMLA.2011.152.

3. Simanjuntak, D. A., Ipung, H. P., Lim, C. & Nugroho, A. S. Text classification techniques used to facilitate cyber terrorism investigation. In *Proceedings - 2010 2nd International Conference on Advances in Computing, Control and Telecommunication Technologies, ACT 2010*, 198–200 (2010). DOI 10.1109/ACT.2010.40.

4. Potha, N. & Maragoudakis, M. Cyberbullying Detection using Time Series Modeling. In *2014 IEEE International Conference on Data Mining Workshop*, 373–382 (2014). URL http://ieeexplore.ieee.org/document/7022621/. DOI 10.1109/ICDMW.2014.170.

Acknowledgements

We thank Conversation AI for making their dataset publicly available.