CS224N Project Proposal

Michael Baumer and Anthony Ho

February 2018

1. **Team:** Michael Baumer (mbaumer) and Anthony Ho (ahho)
2. **Mentor:** N/A
3. **Problem Description:** Many websites with user-submitted content must deal with toxic or abusive comments, but require increasingly fine-grained comment classifications to maintain civility without interfering with normal discourse. In this project, we will improve the identification and fine-grained classification of toxic online comments (Kaggle challenge suggested on course website, online at:

   `https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge`)
4. **Data:** We will be using a dataset of 159,571 comments from Wikipedia's talk page edits which have been labeled by human raters for toxic behavior. The types of toxicity are: `toxic`, `severe_toxic`, `obscene`, `threat`, `insult`, and `identity_hate`.
5. **Methodology/Algorithm:** We will first establish the classification baseline with a simple multi-label logistic regression and averaging all word embeddings of a comment. We will then move on to more elaborate methods like: (1) multilayer perceptrons with word embeddings, (2) RNNs/LSTMs with word embeddings, (3) convolutional neural networks (CNN) with word embeddings and/or character embeddings, and (4) attention-based networks [1]. We plan to utilize existing implementation when available, and to implement our own custom network architectures using TensorFlow.
6. **Related Work:** The earliest work on machine learning-based detection of online harassment can be traced back to Yin *et al.* [2], where they applied SVM on content, sentiment and contextual features to classify harassment. More recently, a paper by the group that has published the Kaggle challenge [3] focused on binary identification of toxic comments (no fine-grained classification). They found success with relatively simple n-gram NLP methods, and left more complex methods like LSTMs as future work. Last year, a group of students looked at the same dataset and found CNNs with character embeddings to be the most successful algorithm [4].
7. **Evaluation Plan:** We will plot ROC curves for classification of each class of toxic comment, and will run on a test set with true labels withheld by Kaggle, scored according to mean column-wise ROC AUC, a statistic which can be computed automatically.
8. **Minimal Requirements:**
   (a) **Number of examples > 10000** ✓
   (b) **Data collection already done?** ✓
   (c) **Feasible task?** ✓
   (d) **Automatic evaluation metric?** ✓
   (e) **NLP required?** ✓

References

[1] A. Vaswani, et al. Attention Is All You Need. arXiv:1706.03762.

[2] D. Yin, Z. Xue, L. Hong, B. Davison, A. Kontostathis, L. Edwards. Detection of Harassment on Web 2.0. In *Proceedings of the Content Analysis in the WEB 2.0 (CAW2.0) Workshop at WWW2009*, 2009

[3] E. Wulczyn, N Thain, L Dixon. Ex Machina: Personal Attacks Seen at Scale. *Proceedings of the 26th International Conference on World Wide Web*, 2017.

[4] T. Chu, K. Jue, and M. Wang. Comment Abuse Classification with Deep Learning. In *CS224n Final Project Report*, 2017.