The Internet Antitoxin: Toxic Comment Classification to Combat Trolls

Zubin Pahuja zpahuja2 | Sarah Schieferstein schfrst2 | Kyo Kim kkim103 | Raghav Gurbaxani raghavg3

22nd February, 2018

Task

Discussing things you care about online can be difficult. The threat of abuse and harassment online means that many people stop expressing themselves and give up on seeking different opinions. Platforms struggle to effectively facilitate conversations, leading many communities to limit or completely shut down user comments.

Existing automated approaches are error-prone, and they dont allow users to select which types of toxicity they are interested in finding (e.g. some platforms may be fine with profanity, but not with other types of toxic content). Our task is to build a multi-headed model thats capable of detecting different types of of toxicity like threats, obscenity, insults, and identity-based hate better than Jigsaw's Perspective API. We are using a dataset of comments from Wikipedia's talk page edits. Improvements to the current model will hopefully help increase participation, quality, and empathy in online conversations at scale.

Going above and beyond the Kaggle contest, we intend to try our model on different relevant datasets with different user demographics, size of dataset, features, structure (such as threaded comments), language and purpose for moderation. One such dataset of interest to tackle the problem of automated moderation is Reddit May 2015 comments dataset, which includes removed comments and the cause for their removal. Another dataset in Greek from Gazzetta sports news portal can be used to assess multi-lingual model.

A potential goal is to build universal models that can generalize well to other platforms to save the cost and time to manually annotate datasets and train task-specific models.

Background

Most commercial systems only use a list of profane words, but the comments such as the example below can easily slip through.

"I sincerely wish that your happy and prosperous life does not last very long!"

Early approaches to comment abuse classification began with Yin, et al., which used an SVM and applied TF-IDF to the features. Recently, deep learning for related fields such as sentiment analysis has proven quite fruitful. RNNs have been known to perform well in sentiment analysis tasks as they use a sequencing model. However, due to the vanishing gradient problem LSTMs are currently the state of the art. Other researchers have used CNNs in sentiment analysis varying from character to sentence level embeddings.

More recently, Straintek, a startup in Greece funded by Googles DNI has published three promising papers in ACL and EMNLP in 2017 as well as demos for multiple platforms on their website StrainTek (2018) .

Data and Evaluation

1. Jigsaw Toxic Comment Classification Challenge on Kaggle (Jigsaw (2018)): contains data from Wikipedia Talk page edits
2. Wikipedia Detox (Wulczyn et al. (2017b)): data from Wikipedia Talk page with annotations for personal attacks (Wulczyn et al. (2017c)), aggression (Wulczyn et al. (2017a)) and toxicity (Thain et al. (2017))
3. Reddit comments dataset (Reddit (2015) and Stuck_In_the_Matrix (2016))
4. Cyber Bullying datasets
   (a) Edwards (2012)
   (b) UWM (2015)
5. Hate Speech (Davidson et al. (2017))
6. Terrorism (AI (2012))
7. Trolling (Golbeck et al. (2017))
8. Gazzetta sports news portal dataset in Greek (Group (2018))

Approach

We intend to use bag-of-words multilayer perceptron, CNN, LSTM, GRU and hybrid or ensemble models with word as well as character embeddings and do comparative analysis on how well the techniques learn and generalize on different datasets, different numbers of training examples etc.

We are also interested in semi-supervised learning and dataless classification for platforms with little or none annotated data.

Kaggle contest platform provides us with an evaluation metric for the Wikipedia Talks dataset. For automatic content moderation, we would evaluate on both probabilistic toxicity score as well as cause for removal. Note: these tasks are different and hence, their evaluation metric will be different as well. For example, the Kaggle contest is a multi-label classification whereas Reddit would be a multi-class classification problem.

To Do

1. Explore the datasets mentioned above and how similar or dissimilar the approach may be
2. Establish a baseline: Logistic regression or keyword based approach.
3. Participate in Kaggle contest. Try different neural architectures such as LSTM or GRU. and engineer our models hyper-parameters for optimal performance on the contest
4. Decide the most efficient evaluation metric.
5. Next, we will explore more open-ended territory such as application to other datasets and the problem of automatic content moderation on larger dataset such as Reddit.

6. We also intend to apply our model trained on a large dataset for classification on a smaller dataset (similar to transfer learning). Or try a dataless or semi-supervised approach.

7. We also intend to apply our model for detection of hate-speech and white nationalism.

Concerns

1. Problems for different platforms can be quite different as the features and labels will be different. So, evaluating a model trained on one dataset on another for generalizability will be difficult.

2. Datasets may not be balanced (down sampling dataset or varying precision recall)

References

UofA AI. The dark web project, 2012. Cyber bullying dataset from wisconsin, `https://ai.arizona.edu/research/dark-web-geo-web`.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee, 2017.

Alexei Bastidas, Edward Dixon, Chris Loo, and John Ryan. Harassment detection: a benchmark on the# hackharassment dataset. *arXiv preprint arXiv:1609.02809*, 2016.

Jayadev Bhaskaran, Amita Kamath, and Suvadip Paul. Disco: Detecting insults in social commentary. 2017.

Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 13–22. ACM, 2017.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515, 2017.

Cicero dos Santos and Maira Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78, 2014.

April Edwards. Chatcoder, 2012. Cyber bullying data which includes texts from MySpace, `http://chatcoder.com/DataDownload`.

Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, et al. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 229–233. ACM, 2017.

Priya Goyal and Gaganpreet Singh Kalra. Peer-to-peer insult detection in online communities. 2013.

Athens University Natural Language Processing Group. Gazzetta sports news portal dataset, 2018. The dataset is in greek, http://nlp.cs.aueb.gr/software.html.

Matthias Hartung, Roman Klinger, Franziska Schmidtke, and Lars Vogel. Identifying right-wing extremism in german twitter profiles: a classification approach. In *International Conference on Applications of Natural Language to Information Systems*, pages 320–325. Springer, 2017.

Lydia-Mai Ho-Dac, Veronika Laippala, Céline Poudat, and Ludovic Tanguy. Exploring wikipedia talk pages for conflict detection. *Book series Translation Studies and Applied Linguistics*, page 146, 2017.

Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*, 2017.

Minlie Huang, Yujie Cao, and Chao Dong. Modeling rich contexts for sentiment classification with lstm. *arXiv preprint arXiv:1605.01478*, 2016.

Jigsaw. Toxic comment classification challenge, 2018. Kaggle toxic comment competition dataset, https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data.

Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

Adam Maus. Svm approach to forum and comment moderation. *Class Projects for CS*, 2009.

Yashar Mehdad and Joel Tetreault. Do characters abuse more than words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303, 2016.

Alejandro Mosquera, Lamine Aouad, Slawomir Grzonkowski, and Dylan Morss. On detecting messaging abuse in short text messages using linguistic and behavioral patterns. *arXiv preprint arXiv:1408.3934*, 2014.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135, 2017a.

John Pavlopoulos, Prodromos Malakasiotis, Juli Bakagianni, and Ion Androutsopoulos. Improved abusive comment moderation with user embeddings. *arXiv preprint arXiv:1708.03699*, 2017b.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Reddit. May 2015 reddit comments, 2015. URL https://www.kaggle.com/reddit/reddit-comments-may-2015. Recently Reddit released an enormous dataset containing all 1.7 billion of their publicly available comments. The full dataset is an unwieldy 1+ terabyte uncompressed, so we've decided to host a small portion of the comments here for Kagglers to explore., https://www.kaggle.com/reddit/reddit-comments-may-2015.

StrainTek. About, 2018. The greek website funded by Google, http://straintek.wediacloud.net/about/.

Reddit User Stuck_In_the_Matrix. I have every publicly available reddit comment for research. 1.7 billion comments @ 250 gb compressed. any interest in this?, 2016. A collection of data collected via Reddit API, https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/?st=jdz37df0&sh=be1a5090.

Nithum Thain, Lucas Dixon, and Ellery Wulczyn. Wikipedia talk labels: Toxicity, Feb 2017. URL https://figshare.com/articles/Wikipedia_Talk_Labels_Toxicity/4563973/2.

UWM. Data and code for the study of bullying, 2015. Cyber bullying dataset from wisconsin, http://research.cs.wisc.edu/bullying/data.html.

Xin Wang, Yuanchao Liu, SUN Chengjie, Baoxun Wang, and Xiaolong Wang. Predicting polarities of tweets by composing word embeddings with long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1343–1353, 2015.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Wikipedia talk labels: Aggression, Feb 2017a. URL https://figshare.com/articles/Wikipedia_Talk_Labels_Aggression/4267550/5.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Wikipedia talk corpus, Jan 2017b. URL https://figshare.com/articles/Wikipedia_Talk_Corpus/4264973/3.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Wikipedia talk labels: Personal attacks, Feb 2017c. URL https://figshare.com/articles/Wikipedia_Talk_Labels_Personal_Attacks/4054689/6.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399. International World Wide Web Conferences Steering Committee, 2017d.

Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB*, 2:1–7, 2009.