A Novel Methodology for Developing Automatic Harassment Classifiers for Twitter

Ishaan Arora, Julia Guo, Sarah Ita Levitan*, Susan E. McGregor and Julia Hirschberg
Department of Computer Science, Data Science Institute
Columbia University, Hunter College*

Motivation

Female journalists are frequently subject to harassment on social media, many consider leaving the profession or change their beats because of it We therefore focus on this subpopulation for our research. Journalists have expressed desire for better engagement tools (Saridou et. al.). Existing tools (Hoffman-Andrews et al. & Chou et al.) are reactive rather than proactive. Our approach focuses on proactive identification of abusive speech.

Challenges

There is no official dataset for evaluating abusive speech. Keyword detection (racial slurs) makes distinguishing between purposeful attacks and jokes between friends and/or community members is difficult. Target words may also be obfuscated. Abusive speech is ill-defined, making it hard to get good annotator agreement on labels.

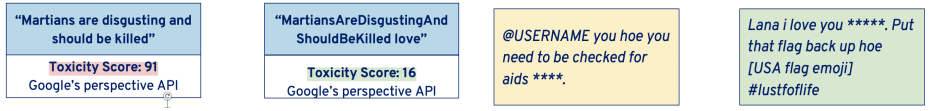

Figure 1. Current Systems are easily fooled

Figure 2. Jocular vs genuine insults

Approach and Data Collection

Direct engagement with a target community for data collection and annotation leads to robust dataset with better labels for contextual and confusing samples. Goal is to use a mixed-methods approach to (A) Construct harassment training data set and (B) Develop classifiers to proactively identify real time abusive speech

Data Collection - Identify abusive tweets using heuristics built from pilot interviews (H1) Tweets from blocked and muted users and (H2) Use Twitter Search API to capture sub-tweeting and snitch-tweeting. Portal present tweets in the context of their larger thread for more accurate annotation. Labels: [hateful, abusive, neutral, spam]

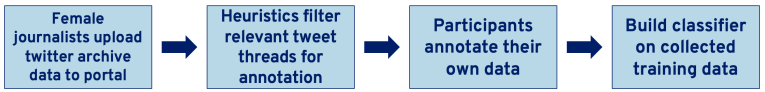

Figure 3. Overview of our approach which directly engages with target community

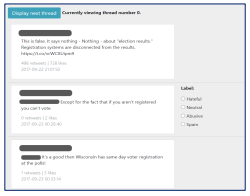

Figure 4. Annotation portal which presents tweet threads for annotation

Modelling Results

Worked with SemEval 2019 Task 5 dataset (HatEval) as Data aligns best with annotated data we expect to gather . Dataset has 9k train records; 1k validation records; test set unreleased.

Task A	Task B
Binary classification task (*hate* | *non-hate*)	Multi-aspect classification task (*hate* | *non-hate*) & (*individual* | *group*) & (*aggressive* | *non-aggressive*)
Winning Approach: Universal Sentence Encoder + SVM (RBF kernel)	**Winning Approach:** Hand-crafted features (Lexical + Syntactic + Bag-of-Words) + SVM (Linear kernel)

Figure 5. Description of sub-tasks for SemEval 2019 dataset

For Task A, we experimented with classical ML and BERT based approach. For BERT, pre-trained BERTForSequenceClassification used. Fine-tuned for ~10 epochs. Significantly improved accuracy and other metrics compared to official task results. For Task B, we used handcrafted features (Lexical + Syntactic + BOW) + SVM (Linear). Could not include LIWC and imperative mood.

Dev accuracy: 75.7%	
Non-hate:	**Hate:**
Precision: 0.815	Precision: 0.695
Recall: 0.747	Recall: 0.773
F1: 0.780	F1: 0.732

Figure 6. Results with BERT based approach

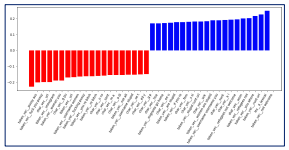

Figure 8. Bag of words features like migrants go home, are among the top features for Task B

Implementation	Accuracy on dev set	Macro-F1 on dev set
Authors	0.65	0.65
Our implementation	0.664	0.665

Figure 7. Replicating wining approach for Task A using classical ML

Task	Author F1	Our F1
Hate / Non-hate	0.71	0.71
Individual / Group	0.87	0.76
Aggressive / Non-aggressive	0.66	0.68

Figure 9. Replicating wining approach for Task B using classical ML

Results and Limitations

Early users shared positive feedback regarding usability of portal. Annotation efficiency: ~300 tweet threads per hour, sufficient for curating a quality dataset in less than 40 hours. Modelling results indicate the robustness of resulting classifier. Limitations: include dependence on archive download functionality of Twitter, which was suspended for 2 months during research period, following July hack.

Data Science Institute
COLUMBIA UNIVERSITY