Final Project Report

Thyne Boonmark, Chloe Lee, Bobby Zheng

July 4, 2020

Abstract

Detection of user generated abusive language or hate speech has become increasingly important in recent academic research, as State-of-the-Art NLP models become more accurate in classification tasks given the boost in contextual word embedding. However, many research and methodologies are primarily focused on a binary identification task of hate speech or some sub-categorical classifications labeled by a group of annotators with subjective or specific guidelines that are more often tied to racial, religious, or ethnicity biases. In this work, we try to identify and analyze languages that can lead to a thread of "inflammatory comments", which themselves can also be inflammatory against members in the community. More specifically, we design a scoring system that captures the overarching concept of "inflammatory" and enables comparisons among comments with varying levels of inflammatory scores. We achieve this with a detailed annotation process and guideline. Lastly, we leverage two NLP modeling strategies to predict inflammatory scores of comments and to analyze the key contribution factors.

1 Introduction and Motivation

The proliferation of inflammatory comments online communities encounter has attracted large amounts of effort from the NLP research community not only because of the divided and hostile environment it could create, but also the waves and magnitude of "a negative psychological impact on [the] target/victim and possibly others who participated in the same conversation"[Mojica de la Vega and Ng, 2018].
In order to contribute to the effort in identifying, predicting, and preventing inflammatory comments, we design a scoring system that captures the overarching concept of "inflammatory" and enables comparisons among comments with varying levels of inflammatory scores.

To approach this, we first defined "inflammatory comments" as comments that tend to excite anger, disorder, or tumult. Then, we quantified comment posts with inflammatory scores based on their associated thread of annotated replies, generated from our annotation process. Lastly, we utilized two NLP modeling strategies to predict the quantified inflammatory scores and extracted relevant features.

More specifically in the modeling process, the prediction outputs are quantitative inflammatory scores where the inputs are either tokenized words from the posted comment or features created from these tokenized words. Additionally, to better understand what language is most effective at inciting inflammatory comment, we performed feature importance and linguistic attention weighting analysis.

2 Related Work

During our research, We noticed an evolution of technique advancements in the space of inflammatory hate speech detection from rule based language processing to deep neural networks. Some of those early studies[Yin et al., 2009], use manually developed regular expression patterns in conjunction with n-grams to identify provocative language. Another approach is to look beyond pattern recognition and use rule based classes syntactic constructs that tend to be insulting or condescending, such as imperative statements[Spertus, 1997]. However, both approaches don't take contextual information into account. [Sood et al., 2012]'s work enhances list-based and rule-based approaches by utilizing a modern Machine Learning methodology (SVM on edit distances) to detect languages that are intentionally misspelled. More importantly, it's one of the first research works that leverages crowd-sourcing (Amazon Mechani-

cal Turk workers) to complete 6,500 classification tasks.

From there, we have noticed an increasing development trend in two closely intervened aspects of research in this space, one that leverages on quick advancement in the NLP models while the other tries to target more reliable and robust annotation process and to build "gold standard" data sets for future researches. For instance, Yahoo researchers in [Nobata et al., 2016] use a wide variety of features including N-grams, linguistic, syntactic (POS) and distributional semantic features (word2vec and comment2vec) into a Vowpal Wabbit's regression model for multi-class classification. In the recent developments such as [Paetzold et al., 2019], deep neural network models are implemented with the aim to achieve better classification performances in multiple tasks or in different languages. Furthermore, in order to capture contextual information, attempts such as [Kudugunta and Ferrara, 2018] makes more complicated modeling by incorporating user metadata as auxiliary input to LSTM structure. We recognize the creativity in all of these attempts but don't believe the focus of our study is simply targeting higher prediction performance on a given data set.

On the other hand, in terms of building "gold standard" data sets, researchers have put in great effort to continuously redefine the annotation process, guidelines, and evaluation metrics to make more robust data sets for the general research community. For instance, in [Golbeck et al., 2017] research, 35,000 tweet comments are annotated by a group of paired annotators who have gone through detailed training with specific annotation guidelines, resulting in a large data set with Cohen's Kappa of 0.84. Furthermore, in [Founta et al., 2018] a comprehensive approach is implemented into data pre-processing, random sampling of data for annotations, and annotation collection process to ensure an unbiased data set and to optimize research cost. This 8-month study goes through several iterations of annotation with up to 20 annotators and provides a groundbreaking data set of 80,000 tweets with majority agreements on 92.5% of the annotations. However, as more resources and efforts are contributed to this space, we notice a general trend of attempts to classify more subtle sub-categories instead of a focus on degree or magnitude of these inflammatory comments.

Different from the previously mentioned annotated labels, our study takes a novel quantitative approach on labeling a comment by aggregating sub-level binary labels for its reply data from Reddit. To our knowledge, similar studies of a quantitative label primarily involve predictions of popularity of comments such as [Klubicka and Fernández, 2018, Zayats and Ostendorf, 2018]. The former uses content, user and tweet specific features while the latter models threaded discussions on social media (Reddit) using a graph-structured bidirectional LSTM. Both are built for predictions in tweet or comment popularity classification tasks. Theoretically different from these approaches, our study focuses on predicting the quantitative inflammatory level of a comment and its likely impact on heated discussions it may create.

3 Data

We pulled our dataset from Reddit using PRAW (python reddit API wrapper). For our annotation, we labelled replies to top level comments (comments made directly to a Reddit post) of various popular Reddit threads from the subreddit r/news.

We used PRAW to collect all comments from the specific Reddit posts that we chose. The comment section of a Reddit post is constructed as a forest, with each top-level comment (comment made directly to the Reddit post) being the head of a tree with each of its direct replies being stored in a list. Each reply is its own instance of a comment forest. The variables we will be collecting from these Reddit comments will be the following:

- **Body_text**: the text of the comment
- **Score**: Reddit's comment score, which is a weighted combination of the number of upvotes and downvotes a comment received
- **Num_reply**: the number of replies a comment received
- **Top 5 replies**: text of the top five replies that the original comment received

As we are most interested in the language of comments that spark online discussions, we focused our analysis on comments that have received at least five replies. In order to manually annotate the discussion surrounding a comment, we

decided to only look at each comment's top five highest scored replies. These replies tend to be the most representative of the discussion, as they have been voted on by most of the users actively engaging with that discussion thread and thus are the replies most likely to be seen.

All of the comment threads we have looked at have been pulled from the top highest scored posts on Reddit's r/news subreddit. Because of these posts' popularity, they have had thousands of comments each, giving us a relatively large sample of comments with replies to analyze. Since we are most interested in the linguistic features of comments and not their content, we specifically sought out Reddit threads where contextual understanding of the topic would not be necessary to identify most inflammatory comments. We avoided topics such as regional politics, sports, etc., since discussion around these topics tend to require a deeper understanding of the subject matter discussed, or are inherently inflammatory. Choosing discussions on more neutral topics helped us avoid noise caused by polarizing opinions, and we are then more likely to capture comments that are inflammatory due to the type of language they use.

Topic posts/Threads used in our project:

- "Japanese firm gives non-smokers extra six days holiday to compensate for cigarette break"
- "Half a billion animals perish in Australian bushfires"
- "Hospitals will have to post prices online starting January 1"
- "YouTube star Daddyofive loses custody of two children featured in 'prank' video"
- "Entire staffs at 3 Sonic locations quit after wages cut to $4/hour plus tips"

4 Methodology

4.1 Construction of Inflammatory Score

The first main part of our project is constructing a quantitative measure of how inflammatory a comment is, which we call an inflammatory score.

Some big challenges to human labeling rise from the existence of so many types of abusive languages and labels that categorize these languages, along with the difficulty in consistently labeling comments; this could lead to considerable disagreement between annotators [Razavi et al., 2010, Founta et al., 2018].This problem becomes much worse once there are more classification categories, compared to binary classification tasks [Nobata et al., 2016].

Given these challenges, in order to create a measure that is more informative than a binary flag yet more straightforward to label with higher inter-annotator agreement, we decided to create a two-fold process where a binary label is assigned to each reply to a top-level comment, and these labels later get aggregated to give each comment a quantitative score.

4.1.1 Data Annotation

For this project, we were the human annotators of the replies to the top-level comments collected from Reddit. We decided to classify replies to comments into binary labels: "inflammatory" or "not inflammatory". We also created a guideline with specific definitions of several different types of languages that would fall into the umbrella of inflammatory language, with an aim to provide a clearer sense of what annotators should be looking out for. The definitions, though condensed into bigger categories due to their similarity, are consistent with those frequently used in existing literature for hate speech or abusive language detection. If the reply falls into any one (or multiple) of the defined categories, annotators will mark it as "inflammatory". Because it is hard to have a single rigid standard for all purposes (or topics, in our project)[Razavi et al., 2010], we also allowed some room for annotators' discretionary judgment to tag replies that are clearly inflammatory in the topic specific context.

- **Offensive Language**: Profanity, strongly impolite, rude or vulgar language expressed with fighting or hurtful words in order to attack or insult a targeted individual or group [Founta et al., 2018, Chen et al., 2012, Razavi et al., 2010].
- **Abusive Language/Mockery**: Strongly impolite language used to abuse, embarrass or show debasement of the reader using crude or provocative language, taboo words, profanity, or unrefined language [Founta et al., 2018, Papegnies et al., 2017,

Park and Fung, 2017, Nobata et al., 2016, Razavi et al., 2010].

- **Hate Speech**: Language used to express hatred towards a targeted individual or group, or is intended to be derogatory, to humiliate, or to insult the members of the group, on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender [Founta et al., 2018, Davidson et al., 2017, Djuric et al., 2015, Badjatiya et al., 2017, Warner and Hirschberg, 2012, Schmidt and Wiegand, 2017].

- **Topic Specific Inflammatory Language** (if applicable): Context specific, topical provocations that are explicitly and clearly intended to incite anger, disorder, or tumult.

In order to create a more robust annotation system that doesn't only depend on one person's judgement, we assigned at least 2 annotators per reply and evaluated inter-annotator agreement by calculating Cohen's Kappa. Disagreements would be resolved by the third annotator(similarly to [Samghabadi et al., 2017]). Ideally, with more resources, we'll be able to annotate more (not just the top 5) replies to a comment and have more annotators assigned to classifying the same reply (which would allow us to decide on a final label based on a majority vote).

4.1.2 Annotated Data Example

Here is an example of a top-level comment and its top 5 replies in the thread: "Half a billion animals perish in Australian bushfires". We have also included how we would annotate each reply, according to the annotation guideline.

– 10 billion land animals are killed annually for food. Why y'all care now?

- It's actually more like 55 billion. "[55 billion land and sea animals die annually to support the U.S. food supply](https://animalclock.org/#section-considerations)" ***not inflammatory***

- At least those died for a purpose. ***not inflammatory***

- *** you, that's why. Also, they *** burned to death. Take your hippy *** elsewhere, you massive stinky ***. ***inflammatory***

- Lol. They don't care. They just pretend they do. **** hypocrites. ***inflammatory***

- 10 billion? You have no clue what you're talking about. More than 72 billion are slaughtered yearly worldwide, not including sea animals. ***not inflammatory***

For this example comment, two of the five replies are labeled as inflammatory.

4.1.3 Label Aggregation

After the annotation process, each top-level comment in our dataset had 5 associated replies that have been classified as "inflammatory" or "not inflammatory". We then assigned each "inflammatory" reply a value of 1 and calculated a comment-level aggregate score that ranges from 0 to 5. This quantitative score is our measure of how inflammatory the top-level comment is, and this score is used in the next portion of this project to build a model that successfully predicts this score when given an arbitrary comment, and to extract useful linguistic features that help with this prediction process. Ideally, with more resources, we should be able to create an aggregate score that doesn't just range from 0 to 5, but actually sums up all the inflammatory replies a certain comment elicits.

4.2 Model Building

There are two modeling strategies we implemented with the main goal of extracting important NLP features that form inflammatory comments:

The first approach was to use a regression or random forest model with self-constructed features in two main categories such as:

- Sentiment features: positive, negative, and neutral sentiments of comments using NLTK's vader sentiment analysis

- Word usage features: We used word usage features based on prior research done on inflammatory online comments

To train our regression models, we first had to deal with the unbalanced nature of our annotated

comment data, as well as the small size of our total dataset (120 scored Reddit comments). During the annotation process we found that very few comments had an inflammatory score greater than 0. To improve our data before model training, we used upsampling to increase the number of comments with scores greater than 0.

Once we had our data ready we trained two types of regression models: linear regression and random forest regression. These models were specifically chosen since they provide us with some measure of feature importance to evaluate the features we input into the models. Linear regression coefficients and random forest feature important measures were used to evaluate the effectiveness of our features in predicting our inflammatory score.

Many of the features we tested were based on a 1997 rule-based inflammatory comment identification system known as Smokey [Spertus, 1997]. We wanted to verify if the features that were useful in this much older system were still useful today, as discourse on the internet has changed over the years.

In the Smokey system, some of the features that most differentiated between benign and inflammatory comments focused on certain word usages. For example the systemed looked for usage of the words: "thank", "please", and "like" or "love", which were usually indicative of non-inflammatory comments. Usage of negative words in close proximity to "you" meanwhile were strong indicators of inflammatory comments.

Besides these features, we also wanted to test features based on Reddit specific language. During our annotation process it appeared that comments which were edited by their commenter, typically denoted by "EDIT:" being found within the comment, were often found to be inflammatory. We thus included this feature in our model to test if this was the case.

We also wanted to test if different measures of sentence sentiment were also useful in identifying inflammatory comments. To add this feature, we made use of the NLTK vader sentiment package to calculate a positive, negative, and neutral score for each comment.

In addition to the regression model, we utilized a neural network model with an attention layer for a many-to-one prediction task to try an additional method to extract important prediction features. Here, we used a simpler version of how [Chakrabarty et al., 2019] constructed an attention mechanism on their Bidirectional LSTM structure that takes a list of tokens from one comment to predict its associated quantitative inflammatory score. We used 50 dimensional GloVe word embeddings with a vocabulary size of 10k and total RMSE as its loss function. Since this task was meant to simply learn relevant linguistic components of higher scored inflammatory comments instead of producing more accurate prediction scores, we used the entire dataset for training without any adjustments.

5 Analysis

5.1 Annotation and Inflammatory Score

In this project, we each annotated 400 replies over two rounds (200 replies each round). With the annotator overlap (explained in section 4.1.1), we had a binary classification label for 600 replies, which were used to give a quantitative score to 120 top-level Reddit comments. After aggregating these labels, the mean inflammatory score of the comments was 0.383.

Despite explicitly designing our comment scoring system to make the annotation process more straightforward and less controversial, it proved a difficult task to generate consistent classification labels among the three annotators. For the 600 replies, we had a relatively low inter-annotator agreement score (Cohen's Kappa) of 0.28.

One of the key factors that contributed to low annotation agreement was that even with the guideline that we defined, there were many comments that called for the annotator's more subjective value judgement. A large majority of the replies were not as hostile as we had initially anticipated - comments rarely fit clearly into one of the buckets that were defined by our guideline, so there were times where the annotators were left to decide on the classification threshold. This again proves that, as much as it is important to have high-quality, clear-cut data to have high inter-annotator agreement, it is not easily attained especially when using real world data.

Another factor was that there wasn't a clear agreement among the annotators about whether aggravation towards the post content that isn't directly tied with the comment-writer or the partic-

ular subreddit community counts as inflammatory or not. For example, for one of the threads that we collected comments and replies from: "YouTube star Daddyofive loses custody of two children featured in 'prank' video", there was some disagreement among annotators on whether replies that were aggressive against "Daddyofive" (subject matter of the Reddit thread, not the comment-writer or anyone interacting in the thread) should be classified as inflammatory or not.

Our findings leave room for future work; given that there are considerable grey areas when it comes to annotating actual comments, it would help to have a more clearly defined guideline that also includes examples that may be harder to classify. Also, because our main focus is to see what comments incite inflammatory responses, it would be helpful to guide annotators to focus on reactions that are aimed more towards the comment-writer or the target audience in the thread, rather than the thread content. A clearer and more specific guideline and better annotator training (especially if we are to expand this project and have more annotators) could considerably increase the quality of the annotation results.

5.2 Linguistic Features

To evaluate the effectiveness of our models we used Mean Squared Error (MSE) over the test set and compared this to a baseline MSE which was calculated using the average inflammatory score of the training set as predictions on the test set. After comparing our models' predictions on a test set to the baseline's MSE of 12.7, we found that the linear regression model performed slightly worse with MSE 13.45 but the random forest model outperformed the baseline with an MSE of 11.93.

What this tells us is that there is potential in using linguistic features to predict how inflammatory a comment is. The features with high and low importance in our models are shown in Table 1.

From the list of important model features, we can see that the sentiment score features were all helpful in determining the inflammatory score of a comment. Initially, we expected the number of swear words in a comment to not be a good predictor, as swear words were not found to differentiate between inflammatory and other comments in previous research [Spertus, 1997]. The number of times that "you" was used within a comment also appeared to be a good predictor. This could be because comments that are more targeted towards an individual or group may engender more heated responses than a comment aimed at a broader group. If so, in the future it may be worth integrating more coreference features in the model and somehow use the target audience of a given comment as a predictor to its inflammatory score.

Important Features	**Unimportant Features**
Pos. Sentiment Score	Contains "thank" word
Neg. Sentiment Score	Contains "please" word
Neut. Sentiment Score	Contains laughter
"you" Frequency	Neg. word near "you"
Swearing Frequency	
Has been edited	
Has "like" or "love"	

Table 1: List of Important and Unimportant Model Features

Among unimportant features in our regression models, we found that around half of the Smokey inspired features were included. These features, listed in Table 1, were initially expected to be useful as they were able to differentiate between inflammatory and noninflammatory comments in Smokey [Spertus, 1997]. This could be a result of internet discourse changing since 1997, and that is is now common for words like"thank" and "please" to be used in not only positive contexts.

However, it is important to keep in mind that these results are preliminary as we did not have enough data to train well-rounded models and perform extensive evaluation on features. With enough data, we might find that these unimportant features may actually turn out to be good predictors of inflammatory comments but within our sample of comments they appear not to be.

In the neural model, we extracted relative word attention weightings within each comment for all inputs and built a few attention weighting heatmaps for some of the highly inflammatory comments. As an example, Figure 1 demonstrats and verifies some of our earlier observations that pronouns that target a community such as "You" can be a contribution factor for high inflammatory scores. However, since these attention weightings are context and comment specific instead of a being a generalizable feature, other important factors found in earlier models such as "edited comment" weren't particularly helpful in this example.

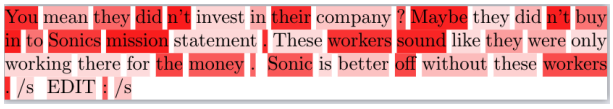

Figure 1: Highly Inflammatory Comment Heatmap Example

5.3 Inflammatory Score Prediction

Using the regression models we made, we now have a way to automatically predict how inflammatory arbitrary comments are. For example, when we use our random forest model to predict the score of the following comment:

- "your opinion sucks, stop posting on this site ****"

The model scores this with a 1.43. Meanwhile if we take a more benign comment like:

- "thats so cool! thanks for posting"

the model gives this comment a lower score of 0.32. These scores make intuitive sense in that the first comment should receive a higher score than the second comment, but due to our small data size the model can only make predictions within a limited range of scores. The lack of many comments with scores greater than 1 limit the possible predictions our model makes. In the future, we would need a much more expansive training set to get better performing models.

6 Conclusion

As the amount of online user generated content grows rapidly, the need to proactively identify inflammatory comments, which could negatively impact online communities, continues to increase. While there has been much work in this area, many are focused on binary or multi-class classification tasks for an individual post or comments. To date, not much research has focused on a quantifiable inflammatory level that can be used to compare incendiary comments. In our work, we defined a methodology to convert this classification task into a regression task. Additionally, we conducted a proof-of-concept trial to demonstrate the validity of this approach with linguistic feature analysis.

Admittedly, there are many areas future work can improve such as robust training for the annotation process at a large scale to obtain a high-quality inflammatory score dataset, additional linguistic features analysis such as syntactic structures impact on inflammatory scores, and semantic meanings analysis as we noticed in the annotation process that context of the comment and its associated topic can be an important factor to identify inflammatory languages.

References

[Badjatiya et al., 2017] Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760.

[Chakrabarty et al., 2019] Chakrabarty, T., Gupta, K., and Muresan, S. (2019). Pay "attention" to your context when classifying abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 70–79, Florence, Italy. Association for Computational Linguistics.

[Chen et al., 2012] Chen, Y., Zhou, Y., Zhu, S., and Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80.

[Davidson et al., 2017] Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

[Djuric et al., 2015] Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., and Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30.

[Founta et al., 2018] Founta, A.-M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., and Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior.

[Golbeck et al., 2017] Golbeck, J., Gnanasekaran, R., Gunasekaran, R., Hoffman, K., Hottle, J., Jienjitlert, V., Khare, S., Lau, R., Martindale, M., Naik, S., Nixon, H., Ashktorab, Z., Ramachandran, P., Rogers, K., Rogers, L., Sarin, M., Shahane, G., Thanki, J., Vengataraman, P., and Gergory, Q. (2017). A large labeled corpus for online harassment research. pages 229–233.

[Klubicka and Fernández, 2018] Klubicka, F. and Fernández, R. (2018). Examining a hate speech corpus for hate speech detection and popularity prediction. *CoRR*, abs/1805.04661.

[Kudugunta and Ferrara, 2018] Kudugunta, S. and Ferrara, E. (2018). Deep neural networks for bot detection. *CoRR*, abs/1802.04289.

[Mojica de la Vega and Ng, 2018] Mojica de la Vega, L. G. and Ng, V. (2018). Modeling trolling in social media conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

[Nobata et al., 2016] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 145–153, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

[Paetzold et al., 2019] Paetzold, G. H., Zampieri, M., and Malmasi, S. (2019). UTFPR at SemEval-2019 task 5: Hate speech identification with recurrent neural networks. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 519–523, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

[Papegnies et al., 2017] Papegnies, E., Labatut, V., Dufour, R., and Linarès, G. (2017). Detection of abusive messages in an on-line community. In *CORIA*.

[Park and Fung, 2017] Park, J. H. and Fung, P. (2017). One-step and two-step classification for abusive language detection on twitter. *CoRR*, abs/1706.01206.

[Razavi et al., 2010] Razavi, A. H., Inkpen, D., Uritsky, S., and Matwin, S. (2010). Offensive language detection using multi-level classification. In *Canadian Conference on Artificial Intelligence*, pages 16–27. Springer.

[Samghabadi et al., 2017] Samghabadi, N. S., Maharjan, S., Sprague, A., Diaz-Sprague, R., and Solorio, T. (2017). Detecting nastiness in social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 63–72.

[Schmidt and Wiegand, 2017] Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

[Sood et al., 2012] Sood, S., Antin, J., and Churchill, E. (2012). Using crowdsourcing to improve profanity detection. In *Wisdom of the Crowd - Papers from the AAAI Spring Symposium*, AAAI Spring Symposium - Technical Report, pages 69–74. 2012 AAAI Spring Symposium ; Conference date: 26-03-2012 Through 28-03-2012.

[Spertus, 1997] Spertus, E. (1997). Smokey: Automatic recognition of hostile messages. In *Proceedings of the 14th National Conference on Artificial Intelligence and 9th Innovative Applications of Artificial Intelligence Conference (AAAI-97/IAAI-97)*, pages 1058–1065, Menlo Park. AAAI Press.

[Warner and Hirschberg, 2012] Warner, W. and Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.

[Yin et al., 2009] Yin, D., Xue, Z., Hong, L., Davison, B. D., and Edwards, L. (2009). Detection of harassment on web 2.0.

[Zayats and Ostendorf, 2018] Zayats, V. and Ostendorf, M. (2018). Conversation modeling on Reddit using a graph-structured LSTM. *Transactions of the Association for Computational Linguistics*, 6:121–132.