Analyzing Harassment on Twitter

Janith Weerasinghe
janith@nyu.edu

ACM Reference Format:
Janith Weerasinghe. 2020. Analyzing Harassment on Twitter. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

With the relative anonymity offered by social media platforms such as Twitter, some users abuse and harass other users on the platform. While social media platforms have taken multiple steps to curb off harassment, it is still a widespread problem. In this paper, we analyze the behavior of online harassers and the accounts they target, and quantitatively answer questions such as: at what rate do individuals become targets of harassment, who are the harassers, how is their behaviour and demographics different from a random user, and do they use newly created accounts?

The first step of our approach was to create a classifier that is able to detect harassing tweets. To do this, we combined the approaches of several previous studies and datasets. Then, using these classifiers, we analyzed the tweets from users included in crowd sourced block lists containing usernames of alleged harassers. We found that their behaviour, in terms of harassment were no different to that of a random user. While this was not the result we expected initially, we believe that this could be because the block-lists that we obtained are somewhat old and the user accounts that engage in repeated harassment maybe suspended by Twitter.

Then we analyzed tweets that mention female politicians and journalists and found that 5%-10% of these tweets are classified as abusive, toxic and hateful. Then we analyzed the user accounts that tweeted such toxic tweets to identify which aspects of these users were different from a random user and which user attributes are associated with higher levels of toxic behaviour. Our results show that most of these users are male, and among them, users with newly created accounts, lower tweet frequencies, and lower ages tweeted a higher number of toxic tweets.

2 RELATED WORK

Over the past years, several datasets and classification approaches have been proposed to detect online harassment. The largest such endeavour was the Troll Patrol project [2, 6], lead by Amnesty International. They analyzed tweets mentioning 778 female politicians and journalists in the US and UK and found that 7% of these tweets were abusive or problematic. Furthermore, they found that women of color were 34% more likely to be mentioned in abusive or problematic tweets than white women. This study received wide press coverage and highlighted the scale of harassment on Twitter. While the authors mention that they have plans to release this dataset for the academic community, so far the dataset had not bean made available. Several studies have constructed datasets by sampling Tweets based on a set of harassment related keywords and then manually labelling the tweets, either by crowd sourcing or by trained annotators. Chatzakou et al. [4] collected a dataset of 1.6 million Tweets. They labelled users in to four classes, aggressor, bully, spammer, and normal user. They used textual, network and user features to build a classifier with performance over 90% AUC. Founta et al. [7] created a dataset by sampling tweets that contain at least one offensive words and has a negative sentiment. Then they used crowd sourcing to label these tweets in to four classes: abusive, hateful, spam, and normal. Here, an abusive tweet is one that contains "Any strongly impolite, rude or hurtful language using profanity..." and a hate-speech tweet is one that uses language 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...". Davidson et al. [5] used a similar approach and labelled tweets according to three classes: hate, offensive, and neither. Davidson et al. points out that a major issue in this type of automated classification is that it is often hard to distinguish hateful or harassing tweets from tweets that contain offensive language. Similarly, Golbeck et al. [10] collected a dataset based on keywords related to alt-right, white nationalist, racist, and homophobic content. They iteratively developed a code-book using grounded theory to label tweets in to two classes: harassment and normal. A key difference of this dataset is that the normal class contains words and phrases that usually appear in harassing tweets, but the tweet itsself is not harassing. (E.g. "For the record, I see tax havens as the next world war, not fucking Muslims.")

Another potential avenue for creating datasets is using crowd-sourced lists of Twitter accounts that engage in repeated bad behaviours. Over the last several years, several such automated tools have become available. These tools are usually developed by volunteers. The users of these tools can create a list of users that they would like to block on Twitter. These lists can be published and other users can subscribe to these lists easily. The most widely used platform to host these lists is BlockTogether.org. Geiger [8] presents an extensive summary about these services.

3 DETECTING HARASSING TWEETS

The first stage of our work is to build a classifier that can detect harassing tweets. By using such a classifier, we would be able to identify users that harass users repeatedly, identify their victims and analyze their behaviour and characteristics. To build such a

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

classifier, we used a combination of new features and features described in previous literature, and trained the classifier on several existing datasets.

3.1 Datasets Evaluated

Table 1 summarizes the datasets we used in our work. $\mathcal{D}_1$ is the largest dataset we acquired, which was originally published as a Kaggle competition. Although this dataset contains Wikipedia comments, and not tweets, we still included this dataset due its scale and descriptive labels. Our preliminary tests showed that models trained using this dataset were able to predict tweets accurately. We discussed the datasets $\mathcal{D}_2$, $\mathcal{D}_3$, and $\mathcal{D}_4$ in section 2. $\mathcal{D}_1$ is a multi-label dataset, where each class is labelled independently. Therefore an entry could labelled as both Insult and as Identity Hate. All of the entries that were labelled as Severe Toxic were also labelled as Toxic. The other datasets were multi-class datasets where each entry could only belong to one class.

3.2 Classifier Model

We used a combination of features that are commonly used in natural language classification problems, and some that were used in previous literature. The language use on Twitter is somewhat different when compared to well-formed text. Therefore when computing these features, whenever possible, we used approaches that were specifically designed for social media text.

- **Word N-Grams**: These features include the term frequency – inverse document frequency (tf-idf) of the phrases used in tweets. Before the tf-idf values are computed, we applied the following pre-processing steps. All mentions of usernames are replaced by "MENTION" and urls are replaces by "URL". Then we use NLTK's TweetTokenizer [1] to tokenize the tweets. TweetTokenizer is a Twitter-aware tokenizer that treats ASCII emojis (such as ;) :-) <3) as a single token and limits the number of repeated characters to two (e.g. converts loool, looooooool to lool). Then each token is reduced to its stem using the PorterStemmer [14]. We then computed the Tf-Idf values for uni-, bi-, and tri-grams of these tokens.
- **Part-of-Speech (POS) tag N-Grams**: We used a POS tagger that is trained on Twitter data from the Tweet NLP project [9] that is more accurate at tagging tweets. We then computed the Tf-Idf values for uni-, bi-, and tri-grams of POS tags.
- **Word Clusters:** Word use on Twitter is informal and the same idea, word, or phrase can be expressed in different ways. For example, the tokens *I'll, Ima, imma* and *I'ma* mean the same thing and the words *quite, entirely, particularly, terribly* and *oddly* are semantically related. Clustering such words together and treating them as one token would help in learning generalized language patterns. We use the set of 1000 hierarchical clusters created by Owoputi et al. [13] that are based on English tweets. They computed the clusters using Brown Clustering [3] which assigns words to classes based on the frequency of word co-occurrence resulting in a hierarchical set of classes that are grouped together semantically and syntactically. We replace words in tweets by their cluster identifier and remove words that do not belong to a cluster. We then computed the Tf-Idf values for uni-, bi-, and tri-grams of these clusters.
- **Other Features:** We used the following features that was used by Davidson et al. [5] in their classifier:
  - Quality of each tweet, measured using modified Flesch-Kincaid Grade Level and Flesch Reading Ease scores, where the number of sentences is fixed at one.
  - Sentiment scores computed by VADER, a sentiment analyzer designed for social media text [11]
  - Number of hashtags, mentions, retweets, and URLs, number of characters, words, and syllables in each tweet.

3.3 Classifier Performance

In our preliminary analyses, we tested the above feature sets with Scikit Learn's Random Forest classifier, SVM classifier with a linear kernel, and Gradient Boosting Classifier (GBoost) on the four datasets. Since the performance of the GBoost classifier was the best on almost all datasets, we used this classifier in our future experiments. Table 2 shows the performance of the Gradient Boosting Classifier on the four datasets.

For datasets $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ the classifier performed well at predicting generally abusive tweets (Toxic in $\mathcal{D}_1$, Abusive in $\mathcal{D}_2$, and Offensive in $\mathcal{D}_3$), but did not perform as well in predicting hateful tweets (in $\mathcal{D}_2$ and $\mathcal{D}_3$) and severe toxic comments in $\mathcal{D}_1$. Upon inspecting confusion matrices of these predictions, it became clear that in most cases, hateful and sever toxic entries were misclassified in to abusive, offensive or toxic classes. This is a common problem that was observed by the other similar classification approaches that we discussed in Section 2. Similarly, the low performance of $\mathcal{D}_4$ can be attributed to the same reason since the negative class in $\mathcal{D}_4$ contains tweets that contain offensive words, but are not hateful or rise to the level of harassment.

4 ANALYZING HARASSERS AND VICTIMS

Next, we used the models trained in the previous step to identify users who post harassing tweets (Harassers) and users who are mentioned in harassing tweets (Victims). To identify potential harassers, we first turned towards crowd sourced block lists.

Over the past years, BlockTogether has become the most widely used tool to implement block lists. It allows users to maintain a list of users that they would like to block on Twitter. They also have the option of sharing the url with other people by sharing a randomized URL which points to their block list. Other users can then subscribe to these block lists. Previous tools such as BlockBot and GGAutoBlocker that implemented similar functionality now have have ported their lists to BlockTogether. While this tool was initially developed to ease blocking harassers on Twitter, people also use it to non-harassing accounts such as advertisers and individuals with opposing political opinions. After reading numerous blog posts, news articles, and tweets, we identified three BlockTogether blocklists that have become popular among Twitter users. The first is the level 1 blocklist of BlockBot[1]. This list contains user accounts that "...in the opinion of the blockers the vast majority of our subscribers would likely wish to ignore.". The second is the

[1] http://theblockbot.com

Dataset	Classes	Size	Collection Method
$\mathcal{D}_1$: Jigsaw - PerspectiveAPI [12] Kaggle competition: Toxic Comments	Toxic (9.6%), Severe toxic (1.0%), Obscene (5.3%), Threat (0.3%), Insult (4.9%), Identity hate (0.9%)	159,571	Wikipedia's talk page edits labeled by human raters for toxic behavior.
$\mathcal{D}_2$: Founta et al. [7]	Abusive (27.2%), Hateful (5.0%), Spam (14.0%), Normal (53.9%)	99,996	Randomly sample Tweets and select negative sentiment tweets with at-least one offensive word as positives. Crowd sourced labels.
$\mathcal{D}_3$: Davidson et al. [5]	Hate (5.8%), Offensive (77.43%), Neither (16.8%)	24,783	Sample Tweets containing hate words and crowd source labels.
$\mathcal{D}_4$: Golbeck et al. [10]	Harassment (26%), Normal (74%)	20,360	Sample Tweets containing curated list of derogatory terms and labelled by trained coders.

Table 1: Datasets used in our work, number of records in the dataset, the classes available in the dataset along with the fraction of records belonging to each class, and the dataset collection method.

Label	Precision	Recall	F1	AUC
$\mathcal{D}_1$: *Jigsaw - PerspectiveAPI [12] Kaggle Dataset*				
Toxic	0.90	0.56	0.69	0.95
Sever Toxic	0.51	0.25	0.34	0.98
Obscene	0.87	0.67	0.76	0.97
Threat	0.36	0.18	0.24	0.83
Insult	0.78	0.56	0.65	0.96
Identity Hate	0.59	0.33	0.42	0.96
$\mathcal{D}_2$: *Founta et al. [7] Hate & Abuse Tweets Dataset*				
Abusive	0.89	0.92	0.90	0.97
Hateful	0.70	0.23	0.34	0.84
$\mathcal{D}_3$: *Davidson et al. [5] Hate & Offensive Tweets Dataset*				
Offensive	0.94	0.94	0.94	0.94
Hate	0.60	0.13	0.21	0.86
$\mathcal{D}_4$: *Golbeck et al. [10] Harassment vs. Normal Dataset*				
Harassment	0.62	0.19	0.29	0.65

Table 2: Performance of our classifier on the four datasets. The performance metrics were calculated based on a randomized train (80%) and test (20%) split of the datasets. Here we report the performance metrics for the positive classes. For datasets with multiple classes ($\mathcal{D}_2$, $\mathcal{D}_3$, and $\mathcal{D}_4$), we converted the datasets to a multi-label dataset by "binarizing" the labels.

NaziBlocker[2] blocklist, which is a "...curated block list of fascists, white supremacists, hate accounts etc.". The third is users from the block list GGAutoBlocker[3], intended to block users associated with Gamer Gate. As a control list, we used Twitter's streaming API to randomly sample English tweets and selected the users that made those tweets who had more than 3000 tweets.

Then, we ran our models on the tweets collected from users belonging to each of the above lists. We can use the class probabilities predicted by our classifier as a measure of the "badness" of each of the user tweets. The average of these probabilities can be used as a measure of the "badness" of the user. More specifically, given a model M, a set of user tweets T, and a prediction function $\texttt{predict\_proba}(M, t)$ for $t \in T$ which outputs the class probability, the average score S for a user can be computed as:

$$S(T, M) = \frac{1}{|T|} \sum_{t \in T} \texttt{predict\_proba}(M, t)$$

We will refer to this score as the "harassment score" in proceeding sections. Figure 1 shows the distribution of the harassment scores of the users in each of the block lists and the random list, computed using our different classifier models. Our initial hypothesis was that we would see significantly high harassment scores for the users in the block lists when compared to the users in the random list. However, we observed that the distribution of the harassment scores were similar between the users from the block lists and random users. We believe this could be due to the following reasons: these block lists seems to be created 2-3 years prior to our data collection. Therefore the users might not be active on Twitter, or they may not have posted abusive Tweets recently, or they maybe suspended by Twitter.

As a second experiment, we identified a set of users, who are likely to be targetted for harassment by individuals. Following previous literature [6], we selected a list of female Senate members and journalists who has a follower count between 10,000 and 1 million. We acquired a list of Twitter accounts belonging to Senators from a list maintained by C-Span[4] and identified female senators from this list. Similarly, we acquired Twitter accounts of female journalists from The New York Times, Britebart, and Daily Mail. Our final list consisted 270 user accounts. We will refer to this list as *Victims* in proceeding sections. Then, we collected the tweets that mention these users and ran the above classifiers on these tweets. For each user, we computed the fraction of tweets that mention them that were classified as positive by our classifiers. Figure 2 shows the distribution of these values. This shows that, among the tweets that mention these users, around 5% of tweets are classifier

[2]https://twitter.com/naziblocker
[3]https://twitter.com/ggautoblocker
[4]https://twitter.com/cspan/lists/senators

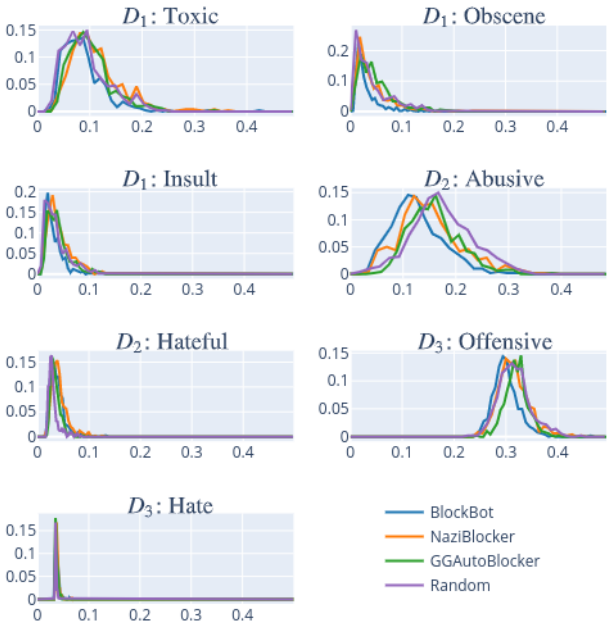

Figure 1: Histograms of the harassment scores for users in each block list, computed under defferent classifier models.

as toxic, abusive, and hateful and around 25% of the tweets were offensive.

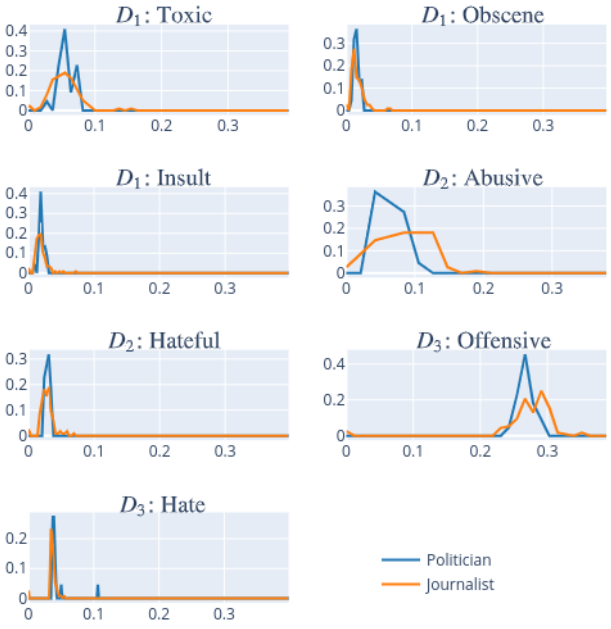

Figure 2: Histograms of the fraction of tweets that mention female politicians and journalists, labelled as positive by our classifiers

We also wanted to analyze twitter users whose tweets were classified as positive by our classifiers in the previous step (we will refer to these users as *Harassers*). To do this we collected their account details and the rest of the tweets posted by these users. For each such user, we computed the fraction of tweets considered as Toxic by our classifier trained on $\mathcal{D}_1$. Similarly, we computed the fraction of toxic tweets for a random group of users. Figure 3 shows the distribution of these values, and shows that the users we identified as harassing the users in our *Victims* list, had a higher fraction of toxic tweets than the random users.

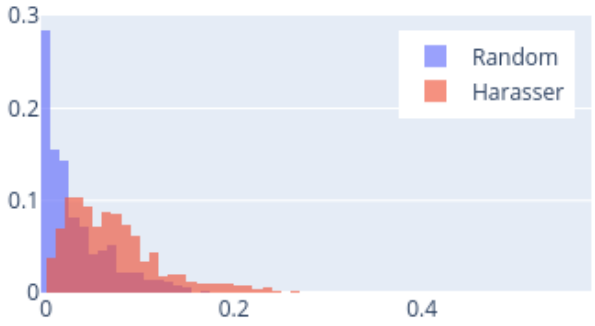

Figure 3: Histograms of the fraction of tweets that were classified as Toxic for *Harassers* and *Random* users.

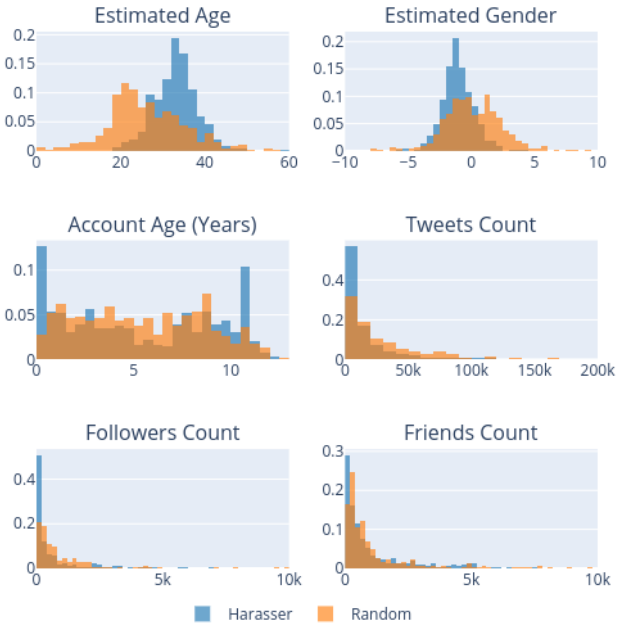

Figure 4: Demographics and account details of *Harassers* and *Random* users. Age and gender are estimated. For gender, a positive value indicates an estimated female user, and a negative value indicates an estimated male user

Next, we analyzed the basic demographics and account characteristics of the users identified as *Harassers*. Twitter does not expose the demographic details such as the age and gender of their users. However, the age and gender of users can be estimated using their language style. We used the demographic classification tool from the World Well-Being Project [15] which has been widely use to estimate demographics of Twitter users. Figure 4 shows these estimated demographic values and the user's account age, tweet count, follower count, and friends fount (the number of users that the given user follows). When compared to random users, the age of the harassers is higher and most of these users are estimated to be male. However, the difference in age can also be attributed to the fact that the users identified as harassers were users who engaged with Senators and journalists, as it is more likely for users of a more mature age to make such interactions.

We wanted to further understand the relationship between the fraction of harassment/toxic tweets that a user has posted, with other demographic and account characteristics that we explored earlier. To examine this, we ran a linear regression with the fraction of harassment tweets as the outcome variable. Table 3 shows the dependent variables that we chose and the linear regression coefficient with the 95% confidence interval. Out of the variables we considered, the user's account age, tweet frequency and their estimated age were predictive of the fraction of harassment tweets that a user made were the ones that were significant at a 5% level. All three variables have a negative coefficient, indicating that users with with a low account age, low tweet frequency, and a lower age had tweeted more harassing tweets. Note the age of a user being negatively correlated with fraction of toxic tweets may seem contradicting to our previous observation that the users in the *Harasser* list had a higher age than a *Random* user. What his result shows is that *among the users* in the *Harasser* list, those with a lower age tweeted a higher number of toxic tweets.

Variable	Regr. coeff. ± 95% CI
Account Age (Years)*	$(-2.40 \pm 1.38) \times 10^{-3}$
Tweets Count	$(1.06 \pm 1.33) \times 10^{-7}$
Avg. Tweet Frequency*	$(-8.24 \pm 6.19) \times 10^{-7}$
Followers Count	$(-1.17 \pm 8.68) \times 10^{-8}$
Friends Count	$(-1.53 \pm 2.28) \times 10^{-6}$
Estimated Gender	$(1.90 \pm 3.36) \times 10^{-3}$
Estimated Age*	$(-1.10 \pm 8.18) \times 10^{-3}$

Table 3: Linear regression coefficients. * indicates statistical significance at 5% level.

5 DISCUSSION

In this work, we created several classifier models that can predict if a given tweet contains language that is harassing, hateful, abusive, and toxic. Using these models, we analyzed tweets of users who are included in crowd-sourced block lists and users who mentioned female politicians and journalists. We found that 5%-10% of these tweets that mention them are classified as abusive, toxic and hateful. We then analyzed various attributes of these users and found that most of these users are male and among them, users with newly created account, lower tweet frequency, and lower age tweeted a higher number of toxic tweets.

While these results are interesting, there are some limitations of our work. First, as is the problem with many automated harassment classification approaches, our classifiers are not perfect and often misclassifies offensive word use as harassment and vice versa. In future work, we would like to analyze the misclassifications made by our classifiers and work on improving their precision and recall. We also limited our analysis to users who targeted female politicians and journalists. Similar analyses could be conducted on other vulnerable groups. Nevertheless, our work further highlights that online harassment is a common problem in social media platforms, and sheds some light on the individuals that harass other social media users.

REFERENCES

[1] [n.d.]. http://www.nltk.org/api/nltk.tokenize.html.

[2] Amnesty International. 2019. Troll Patrol Findings. https://decoders.amnesty.org/projects/troll-patrol/findings

[3] Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based N-gram Models of Natural Language. *Comput. Linguist.* 18, 4 (Dec. 1992), 467–479. http://dl.acm.org/citation.cfm?id=176313.176316

[4] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on Twitter. In *WebSci 2017 - Proceedings of the 2017 ACM Web Science Conference*. 13–22. https://doi.org/10.1145/3091478.3091487 arXiv:1702.06877

[5] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*. 512–515. arXiv:1703.04009 www.facebook.

[6] Laure Delisle, Alfredo Kalaitzis, Krzysztof Majewski, Archy de Berker, Milena Marin, and Julien Cornebise. 2018. Troll Patrol Methodology Note. , 15 pages. https://decoders.azureedge.net/data-viz/images/Troll%20Patrol%20-%20Methodology.pdf

[7] Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *12th International AAAI Conference on Web and Social Media, ICWSM 2018*. 491–500. arXiv:1802.00393 http://arxiv.org/abs/1802.00393

[8] R. Stuart Geiger. 2016. Bot-based collective blocklists in Twitter: the counter-public moderation of harassment in a networked public space. *Information Communication and Society* 19, 6 (jun 2016), 787–803. https://doi.org/10.1080/1369118X.2016.1153700

[9] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech Tagging for Twitter: Annotation, Features, and Experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2 (HLT '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 42–47. http://dl.acm.org/citation.cfm?id=2002736.2002747

[10] Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J Martindale, Shalmali Naik, Heather L Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. 2017. A large human-labeled corpus for online harassment research. In *WebSci 2017 - Proceedings of the 2017 ACM Web Science Conference*. 229–233. https://doi.org/10.1145/3091478.3091509

[11] C. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109

[12] Jigsaw. 2017. Toxic Comment Classification Challenge | Kaggle. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

[13] Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, and Nathan Schneider. 2012. Part-of-speech tagging for Twitter: Word clusters and other advances. (2012).

[14] Martin F Porter. 1980. An algorithm for suffix stripping. *program* 14, 3 (1980), 130–137.

[15] Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Hansen Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1146–1151.