Agent-based user-adapted Filters for Categories of Harassing Communication

Zenefa Rahaman
The University of Tulsa
Tulsa, Oklahoma
zenefa-rahaman@utulsa.edu

Sandip Sen
The University of Tulsa
Tulsa, Oklahoma
sandip@utulsa.edu

ABSTRACT

Online communication and information sharing platforms have witnessed rapid growth in usage leading to peer-to-peer communication at unprecedented scale and diversity. Unfortunately, these platforms also witness an abundance of hateful aggression and harassment towards individuals targeted because of their identities or expressed opinions. Such online harassment can have a significant negative impact on individual health and social relationships. Individuals vary in their perception and sensibility towards different categories of harassing communication. Individual preferences can be used to develop an adaptive agent-based intervention mechanism to protect social media users from harassing communication.

We present the design and implementation of an user-adapted agent-based harassment filter that adapts to user sensibilities and filtering preferences continuously. In this paper, we introduce a taxonomy of harassing communication types, describe the process of collecting and analyzing a crowdsourced Twitter dataset highlighting the need for agent-based user-adapted filters, and present comparative results from user-adapted filters trained with different learning schemes that conform to user filtering preferences. Results show that user-adapted, user-specific filters can significantly outperform general filters even with limited user input

KEYWORDS

Cyber harassment; detection; adaptive agent; agent-based filter;

1 INTRODUCTION

Today's citizens are highly engaged in online social media: both the number of users and the time spent online per user is on the upswing. According to a report by the Pew research center, nearly 76% of American adults use social networking sites in 2017, which was only 7% in 2005 [10]. These social media platforms become valuable repositories of people's opinions and sentiments about services they use as well as their political and religious views. Hence, these sites have key influence on user's opinions and sentiments. Correspondingly, the collected data serve as valuable data sources for businesses, researchers, and policymakers. Whereas these new communication channels, such as online social networks [19] and news sharing sites [20], offer myriad opportunities for knowledge sharing and opinion mobilization [6], they also reveal an abundance of unfortunate intimidatory and hateful aggression [27] towards individuals targeted [37] because of their expressed opinions or identities. Cyber-harassment involves any aggressive and unwanted online communication with the intent to target and intimidate a victim. A recent study [14] found that 40% of adult Internet users have experienced online harassment with young women enduring particularly severe forms of it. 38% of women who had been harassed online reported the experience could be described as extremely upsetting. This victimization of individuals [9] have significant social costs varying from social ostracism to opinion marginalization and suppression, and can cause severe health detriments ranging from anxiety [31] to depression [39] to suicide ideation [18, 29].

Victims of such online attacks are often minority groups or individuals voicing dissent, professionals covering controversial topics essential for informing a democratic society, and groups raising awareness about important issues [4]. A study of offline and online harassment of female journalists found that two-thirds of the respondents reported acts of intimidation, threats, and abuse related to their work [3].

While several computational studies have developed automated mechanisms for detection of unwarranted victimization and harassing attacks on social network and microblogging platforms [26, 32, 40], more comprehensive detection and intervention mechanisms that are grounded in well-founded, interdisciplinary theory of human aggressive and predatory behavior is needed. In this paper, we present a taxonomy of harassment categories to characterize different types of hateful and abusive rhetoric that is common in online social media platforms. This taxonomy will facilitate the development of harassment filters, our research focus as elaborated below.

The goal of this paper is to develop an agent-based user-adapted harassment filters that continuously adapt to individual user's threat perception, tolerance, and sensitivity. As users may have varying sensibility towards different types of harassment, our first goal is to understand the need for agent-based user-adapted filtering of harassing tweets.

To gather ground truth data necessary to identify varying harassment perceptions and to demonstrate the feasibility of training user-adapted filters, we needed user-labeled data-sets. To obtain this data, we first collected a set of ≈ 5230 tweets using the Twitter API and matching keywords associated with the categories identified in the taxonomy. Thereafter, we used a crowdsourcing service, `Amazon Mechanical Turk`, where `MTurk` workers were tasked to label presented tweets based on their perceived harassment intensities and if they wanted to filter them. We collected approximately 26,300 responses from 360 participants for this study (see Section 4 for further details). The goal of the survey is to understand the variation of sensibility among the users, the variation in the tolerance or acceptance of harassment categories in our taxonomy, and to build agent-based user-adapted filter mechanisms. The collected data is used to find the answers to the following research questions:

(1) How user sensibility vary from category to category?

(2) How the perception and the acceptance of the different type of harassment vary in a population?
(3) How the same data can have different impacts on users and how their reactions vary?

We performed an in-depth analysis of the collected Survey Data to identify the variation of harassment sensibility and tolerance levels among the users over different harassment categories in our taxonomy. Our analysis show how users' perception of harassment intensity and filtering preference varies depending on harassment categories. Also, different users' perceived intensity and acceptance can vary significantly for the same tweet. We evaluated filtering mechanisms for each user and found that agent-based, user-adapted filters are more accurate in predicting user preferences when compared with a general filter trained on the entire dataset. The general filter failed to protect a sensitive user from exposure to unwanted or unacceptable tweets as the general filter learns the filtering need of the tweets based on the majority of filter/no-filter labels over the population. The observations support the need for and benefit of user-adapted filtering of harassing communication.

2 CATEGORIES OF CYBER-HARASSMENT

Most cyber-harassment research that we have come across have primarily focused on binary classification of harassment: a communication either contains harassing content or does not [11, 13, 35].

Though a standard definition of online harassment does not exist, most definitions include the following key components: (1) unwanted behavior that occurs through electronically mediated communication, (2) behavior which violates the dignity of a person by creating a hostile, degrading, or offensive environment [5]. Behaviors can include offensive name calling, attempts to embarrass, physical threats, stalking [14], gender harassment, unwanted sexual attention, sexual coercion, denigration (sending harmful or cruel statement about a person to other people online), impersonation (pretending to be another person in order to make that person look bad), flaming (sending angry, vulgar or rude messages about an individual through an online, public forum), and exclusion (the exclusion of an individual from an online group) [33]. Definitions vary in terms of whether the behavior must occur multiple times, intentionally cause harm, and/or involve a perpetrator known to the victim [14].

[25] have emphasized the need for clear definitions in research for the field to progress. Much of the current literature examining offline harassment require knowledge of a perpetrator's intentions which, while difficult to discern in offline environments, is almost impossible to confirm in online environments such as social media platforms. For this reason, online harassment as currently understood by researchers is defined in terms of a victim or third party's understanding rather than a perpetrator's motives. Researchers believe there are three main reasons perpetrators may purposefully engage in harassing behavior. Purposefully harassing behavior includes (1) rude comments used as a form of self-expression [12], (2) intimidation strategically designed to (a) interrupt communication on a topic or (b) retaliate for past reports or comments [4]; and (3) acts with no strategic aims other than causing psychological or physical harm [7]. Further complicating an understanding of harassment is the variability in how harassment is perceived. Many definitions require the victim to view the behavior as offensive or threatening [16]. Understanding a victim's reaction is equally as difficult as discerning a perpetrator's motive when researchers are unable to directly communicate with the victim.

To understand different types of online harassment, a taxonomy was prepared containing different categories of cyber-harassment that was paired with an associated vocabulary. The initial category list is generated from existing literature [14, 28, 33] :(i) Insults and name calling, (ii) Sending harmful or cruel statement about one user to others, (iii) Religious/racial/ethnic epitaphs, (iv) Sexual orientation, (v) Sex/ gender, (vi) Threat, (vii) Multiple type (message contain more than one harassing type), (viii) Revealing personal information online about target, (ix) Tapping, computer viruses, and other digital security threats, (x) General threat of physical harm to self or others, (xi) Specific threats of physical harm with individual, and (xii) Impersonation online.

After analyzing a stream of tweets over a period of time, we observed that some of the categories that we initially listed did not appear on Twitter streams. Accordingly, we paired the initial list down to the following that was used in our subsequent data gathering and experimentation:

General harassment: Communication is harassing, but does not easily fit into any other identified category.

Cruel statement: Communication contains information that is negative and personal. This information is directed at a specific target and does not specifically address a person's religion, race, ethnicity, sexual orientation, or gender/sex.

Religious/racial/ethnic slurs: Communication designed to highlight or attack a person's race, religion, or ethnicity.

Harassment based sexual orientation: Communication designed to highlight or attack a person's sexual orientation.

Sexual harassment(Gender based harassment): Communication contains a short, negative label designed to highlight or attack a person's sex or gender.

Threats of physical harm/violence: Communication contains a direct threat, metaphorical or actual, against a person's property, family, self, or digital presence.

Multiple types: Communication contains more than one type of harassment defined.

Non-harassment: Communication does not fit into any of the discussed harassment categories.

We also identified keywords for each of the identified harassment categories based on frequent words used in corresponding tweets.

3 HARASSMENT DATA SET

We chose the Twitter platform for collecting harassing communication. Using Twitter's streaming API, we obtained a set of tweets matching the keywords associated with the initially identified categories and stored them in a MySql database. As is the case with many real-world data, the collected data had several issues which required cleaning and pre-processing to facilitate further research and analysis. After pre-processing, 5231 out of 8000 collected tweets were found to be usable. For each tweet in this set, we had three individuals label it according to one of the category types in our taxonomy. The majority label was then associated with the tweet and used for the remainder of our experiments reported in this paper.

Table 1: Frequency of Tweet Categories

Category	# of tweets
General harassment	79
Cruel statement	1054
Religious/racial/ethnic	89
Sexual orientation	11
Sex/ gender	656
Threat	236
Multiple types	106
Non-harassment	2382
Non-codable	618
Total data	5231

The result of the labeling is presented in Table 1 which contains the number of tweets assigned to each of the aforementioned categories. We refer to this dataset as the `Twitter Categorical dataset`. Almost 45% of this usable data are categorized as non-harassing or do not fit into any given categories. We observed, that several data points lacked in unanimous labels. This observation shows that there is a pronounced variation in harassment perception among the human labelers. This perception variability results in similar tweets being assigned different category labels.

4 SURVEY DATA: USER-BASED SENSITIVITY DATA

We wanted to understand how different users have distinct sensibilities towards different harassment categories. We used Amazon Mechanical Turk [34] workers to rate harassment intensities and the need for filtering of individual tweets[1]. The MTurk survey participants were provided a collection of unlabeled tweets, drawn from multiple harassment categories of the Twitter Categorical dataset. Participants were asked to respond to the following two questions about each tweet:

(1) What intensity of harassment is present in the following tweet?
Options: *(1) None, (2) Minimal, (3) Moderate, (4) High, (5) Extreme.*
(2) Do you want to filter this tweet?
Options: *(1) Yes, (2) No.*

The survey is prepared using Django, a based web application hosted on the Heroku website [23]. The link of the survey was posted on Amazon Mechanical Turk website to reach MTurkers. For each MTurk participant, the survey contained 75 tweets selected from all categories of the Twitter Categorical data. Each tweet was presented to 5 MTurk workers. Around 360 MTurkers successfully completed this study over a two week period which allowed us to collect around 26,500 responses. We used this data set to understand how perception and acceptance of the different type of harassment vary in a population. As described below, the analysis of the dataset also provided the rationale for user-adapted , agent-based harassment filters learnt from user filtering choices.

[1]Amazon Mechanical Turk (MTurk) is a crowdsourcing platform where researchers can post human intelligence tasks to be performed by paid human workers (MTurkers).

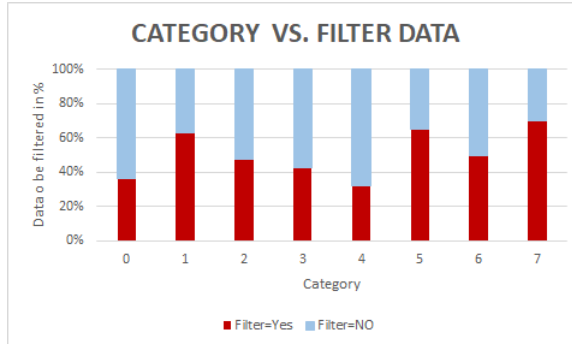


Figure 1: Percentage of data selected by users to be filtered for the different harassment categories.

5 NEED FOR USER-ADAPTED FILTERS

We wanted to understand if different users have distinct harassment perception and filtering sensibility/preference, with the goal of determining if user-adapted agent based harassment filters are warranted.

As mentioned above, each user was asked for each of the tweets presented, if (s)he would prefer that tweet to be automatically filtered from their tweet stream. The percentage of data in each of the harassment categories that the users wanted to be filtered is shown in Figure 1. In Figure 1, the X-axis represents different harassment categories and the Y-axis represents the percentage of tweets selected to be filtered. Each bar represents one category and the 'red' colored part of the bar represents the percentage of tweets to be filtered for that category. We observe there are variations in the percentage of data to be filtered for different categories.

Observation 1. *User filtering preferences or acceptance of cyber-harassment vary by harassment categories.*

Figure 2 depicts the percentage of tweets to be filtered for each category given the perceived harassment intensity. In the figure, the X-axis represents different intensity levels. The Y-axis represents the percentage of tweets selected to be filtered. Each colored line corresponds to a particular harassment category.

Observation 2. *The percentage of tweets selected for filtering increases with the increase in perceived harassment intensity.*

This pattern holds for each of the harassment categories. This observation supports the hypothesis that users have lower acceptance of higher intensity of cyber-harassment. The figure also shows that the tolerance level does vary between categories for the same harassment intensity. That leads to the following critical observation:

Observation 3. *Harassment tolerance depends both on the category of harassment as well as the perceived intensity of harassment.*

5.1 Intensity level influence on filtering choice

In this section, we present results from statistical tests to determine the significance of the observed trends in the survey.

5.1.1 ANOVA test: We used one-way ANOVA tests to determine if the choice to filter tweets varied significantly by the perceived harassment intensity level. ANOVA checks the impact of one or more

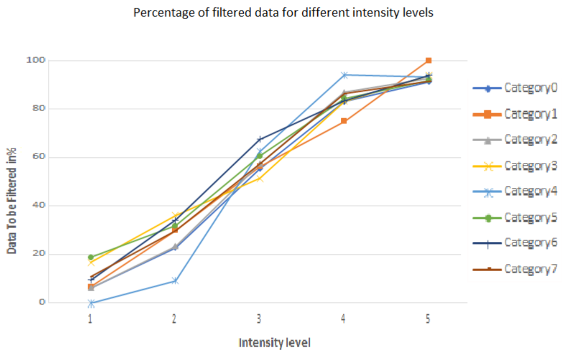


Figure 2: Percentage of tweets selected to be filtered for each category and for different perceived harassment intensity levels. Y-axis represents the percentage of tweets selected to be filtered and the x-axis represents different intensity levels.

factors by comparing the means of different samples. Using this test we can conclude if there is a statistically significant difference between the percentage of tweets selected for filtering between different perceived harassment intensity levels. The following two hypothesis are considered in ANOVA testing:

Null hypothesis (H0): All intensity levels have the similar effect on the population of users

Alternate hypothesis: Different intensity levels have the significantly different effect on users

The ANOVA test for the collected data for different intensity levels yielded an F-statistic value that measures the differences in the means of different intensity levels and suggests whether the levels are significantly different or not. α is considered the significance level and the probability of rejecting the null hypothesis while it is true. If the F-statistic is more than the F-critical value for the chosen α value, then the null hypothesis (H0) can be rejected and we can say that different intensity levels have significant effects on users. For the data presented above, ANOVA test returned an F-value of 304.7577 which is greater than the F-critical value, 2.64146, for the selected alpha level of 0.05. The significance can also be determined by comparing the p-value calculated with the $\alpha = 0.05$ value selected: if the P-value is less than α, we can reject the Null Hypothesis. The ANOVA result is $[F(4, 35) = 304.758, p = 1.139e^{-26} < 0.05]$, and based on calculated p-value ($p = 1.139e^{-26}$). Hence the null hypothesis (H0) can be rejected and we claim that different intensity levels have significantly different effect on user filtering selections.

5.1.2 Effect size: In addition to checking the statistical significance, it is useful to report the effect size measure for an ANOVA test, which reflects the size of the performance difference of the alternatives being considered. A high effect size value signifies that not only there are more than two groups which significantly differ from each other but also suggests that the difference is significantly high. For the ANOVA test we ran, the effect size value, $\eta^2 = \frac{SS_{intensities}}{SS_{Total}}$, is 0.972. This suggests that some of the intensity levels are different from others by a high margin.

The limitations of the one-way ANOVA testing is that while it can suggest that at least two of the tested data groups were different from each other, it cannot identify which groups were different. To identify those groups, we needed to further test what are the intensity levels that result in significant differences.

For this purpose we use Tukey test which is a statistical significance test that is often used to identify the effect size with ANOVA analysis.

5.1.3 Tukey Honest Significant Difference (HSD):. Tukey's Honest Significant Difference (HSD) test or Tukey test is a post-hoc test based on the studentized range distribution. Tukey's HSD is considered to be the strongest test to find out which groups are significantly different and is widely used. This test calculates all possible pairs of means and then calculate the significant difference. Based on the computed q-statistics by Tukey test on our dataset, we observe that Intensity level 1-3 are significantly different from all the other intensity levels (1-5) with a 99% confidence. Intensity 4 is different from intensity 5 with 95% confidence value. All the pairwise comparisons support the alternative hypothesis.

5.2 Influence of Harassment Categories on filtering choice

We have so far presented statistical analysis to understand whether user's filtering choice are affected by the perceived intensity level. In this section, we discuss the statistical significance of user's filtering choice as influenced by the the particular category of harassment contained in the received communication. In particular, for a given perceived harassment intensity level, whether the percentage of tweets marked for filtering varied for different harassment categories? It is possible that such differences can exist for only some, but not for all, perceived intensity levels. As we have only one value, percentage of tweets marked for filtering for each harassment category, for each intensity level, we needed a different significance testing mechanism than ANOVA. In the following, we present the results from using the Confidence Interval Coefficient measure used for this analysis.

5.2.1 Confidence Interval Coefficient: Confidence Interval Coefficient [2] compares proportions of samples through constructing simultaneous confidence intervals and P-difference for the confidence intervals. We constructed confidence intervals to make pairwise comparisons of the percentage of tweets to be filtered for different categories for a given intensity. The significance is determined by comparing the p-difference with the α value (we chose $\alpha = 0.5$). If the p-difference calculated for confidence intervals is greater than the α level selected, it suggests that for the same intensity level there is a significant difference in the chosen categories. In Figure 2, we have presented the percentage of tweets to be filtered for each category and for all the intensity levels. Results of Confidence Interval Coefficient for intensity level '2' is shown in Table 2. The rows that are highlighted in green indicates the categories which are significantly different from each other for the given intensity level. Results in Table 2 suggests that for intensity level '2', users choose a different percentage of tweets for category 0 compared to category 3-7 and these differences are statistically

Table 2: Confidence Interval Coefficient for Intensity Level '2' (columns highlighted denote statistically significant difference; 19 pairs are different).

Category	Category	P-Diff	Lower	Upper
0	1	0.0704	-0.2139	0.0759
0	2	0.0055	-0.1448	0.1339
0	3	0.133	-0.2791	0.0182
0	4	0.137	0.01433	0.2546
0	5	0.0898	-0.2342	0.0582
0	6	0.1142	-0.2597	0.0357
0	7	0.0698	-0.2133	0.0764
1	2	0.0649	-0.0818	0.2091
1	3	0.0626	-0.2157	0.0929
1	4	0.2075	0.0764	0.3307
1	5	0.0194	-0.1709	0.133
1	6	0.0438	-0.1963	0.1105
1	7	0.0006	-0.15007	0.1513
2	3	0.1275	-0.2742	0.02415
2	4	0.1427	0.01909	0.2607
2	5	0.0843	-0.2294	0.0641
2	6	0.1087	-0.2548	0.0416
2	7	0.0632	-0.0923	0.2163
3	4	0.2702	0.1335	0.3963
3	5	0.0432	-0.1132	0.1979
3	6	0.0188	-0.1385	0.1754
3	7	0.0632	-0.0923	0.2163
4	5	0.2269	-0.3511	-0.0938
4	6	0.2514	-0.3767	-0.1162
4	7	0.2069	-0.3299	-0.0758
5	6	0.0244	-0.1786	0.13073
5	7	0.01999	-0.1324	0.17157
6	7	0.0444	-0.1098	0.1969

significant. For space considerations, we are not presenting the tables for other perceived harassment intensities. We have observed that user acceptance of different harassment categories vary significantly at lower intensity levels. As the perceived intensity increases, however, the difference in the acceptance of harassment of different categories faded. This is primarily due to the fact that more and more tweets, irrespective of categories, are selected to be filtered by users at high perceived harassment intensity levels. Observation from the statistical test on the user responses confirms that user choice of tweets to filter is dependent on both the category and the intensity level of a particular tweet. Results from ANOVA testing and Tukey test shows that user filtering choice have significant differences with varying intensity levels. The corresponding effect sizes were also found to be high. Results of Confidence Interval Coefficient suggests that the user reaction varies for certain categories of harassment for the same intensity levels. These observations highlight the need for user-adaptiveness.

6 USER-ADAPT HARASSMENT FILTERS

We now focus on developing agent-based user-adapted filtering of harassment content for the users. As we observed above, users acceptance or tolerance for harassing communication varies based

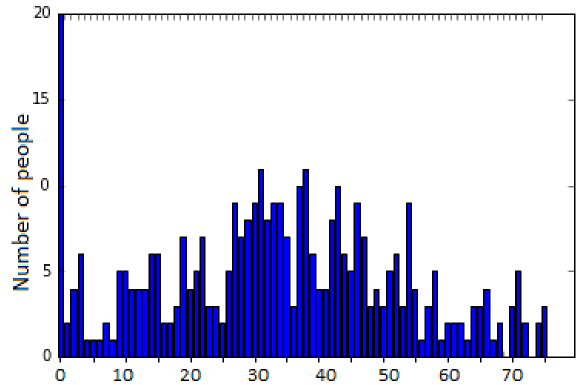
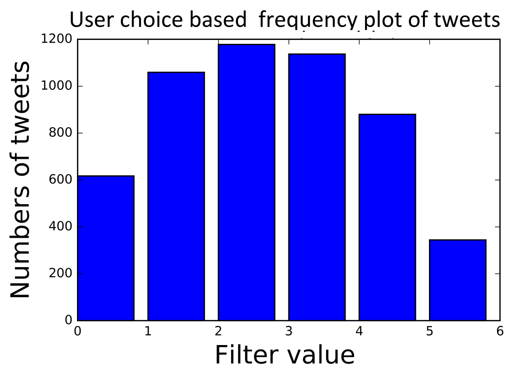


Figure 3: Histogram of # users wanting to filter certain number of tweets. X axis represents # of tweets (0-75), Y axis represents # of users wanting to filter that many tweets out of the 75 presented.

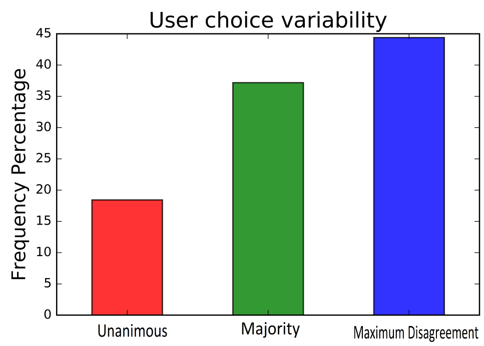
on intensity level and harassment category. In addition, different users have a different filtering preference for different intensity levels and categories. So a general filter, trained on all labeled data, is going to be ineffective to meet the filtering needs of individual users.

To develop an adaptive personal agent for filtering harassing communication, we further analyzed the Survey Data. In Figure 3, we present a histogram of the counts for users with the number of tweets they wanted to be filtered. Each bar in the histogram shows how many people wanted to filter that number of tweets out of the 75 tweets they labeled. The x-axis of the histogram represents the number of tweets (0-75) and the Y-axis represents the number of people who wanted to filter that many tweets out of the 75 tweets presented. The histogram demonstrates that there is considerable variation in the population based on individual user's acceptance of harassment and hence the need for user-adapted filtering of harassing communication. The results clearly demonstrate that there were participants who were extremely tolerant about harassing content and chose not to filter any of the tweets. On the other end, there were participants who were extremely sensitive to harassment and chose to filter all the tweets they receive based on the contents. Variation in sensibility is reflected in the variations in the number of tweets to be filtered as selected by the users.

We recall that each tweet from the Twitter Categorical data was labeled by 5 MTurk users for harassment intensity and filtering preference. In Figure 4 (a) we present a count of tweets that received a certain number of votes, from 0-5, for filtering. In Figure 4 (b) we show how many tweets were unanimously voted by all the five users to filter/not-filter, how many got significant *majority* (4:1 or 1:4 ratio) of filter/not-filter selection, and how many tweets had *maximal disagreement* (3 to 2 or 2 to 3 counts) for filter/no-filter choices. Results show that only 18% of the tweets were unanimously voted on for filter/no-filter. Approximately 38% and 45% of tweets were marked for majority and maximal disagreement groups respectively. These observations clearly demonstrates that different users vary in their harassment sensibility and acceptance.



(a)



(b)

Figure 4: (a) Frequency plot of tweets labeled for filtering by 0-5 users (X-axis represents how many users, out of 5, chose to filter a tweet, Y-axis represents the # of tweets). (b) User Choice variability plot.

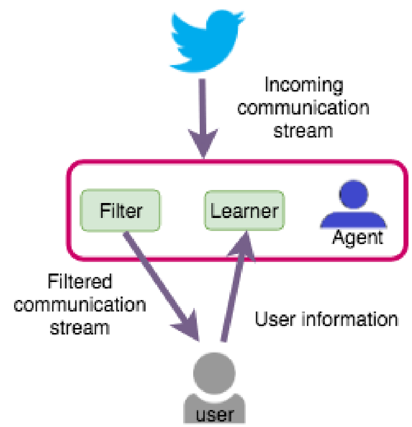


Figure 5: An agent-based user-adapted harassment filter.

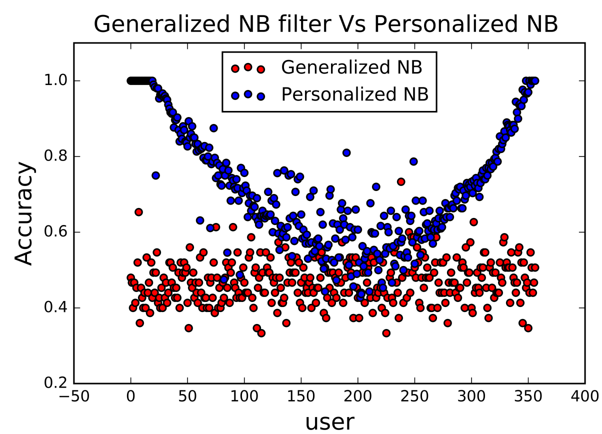


Figure 6: Accuracy of General Vs user-adapted Filters.

6.1 A general harassment filter

Before developing agent-based user-adapted harassment filters, we created a general filter that learns to filter based on the majority of filter/no-filter value for each tweet. The filter-label of each tweet is decided based on the majority of the 5 votes. Developing a single filter for all users, rather than user-adapted, agent-based filters learnt from interactions with each user, is computationally cheap. We believe, however, that a single general filter will not be effective for all users because of significant false positives and false negatives, and hence precision and recall errors. In particular, such a general filter will not be able to protect a sensitive user from exposure to unwanted or unacceptable harassing tweets.

6.2 Agent-based user-adapted filters

In the agent-based filter mechanism, the agents are not only initially trained by learning on the user filtering choices but can also continuously monitor and adapt based on the users' online behavior. Based on these interaction experience, agents can adapt their filtering behavior to best reflect changing user sensibilities and protect them from harassing encounters on their social media accounts. In our experiments, we allocated one agent for each user in the system (Figure 5 presents the interaction between the agent and its associated user.). We provide training data from the filtering choices on tweets to an agent from its associated user. That agent then used effective supervised classification algorithms to learn user-adapted filters for its user. With experience an agent learns about user's different interaction preferences, and classifiers trained using this gathered knowledge helps the agent adapt to current user preferences. While this modality of agent-based filter development requires much more effort, as one agent has to be trained for each user, we believe that performance improvements in intervention by shielding users from unwanted harassing communication, justifies this additional computational cost.

We used Naive Bayes (NB) [22], Support Vector Machine (SVM) [17] and Random forest classifier [21] to train user-adapted filters. We also use the same classification mechanisms to build the general filters. In Figure 6, we present a comparison of the 10-fold cross-validation accuracy of a Naive Bayes based user-adapted filter for each agent to that of the corresponding general filter when used for each user. In Figure 6, the X-axis represent each user and the Y-axis represents the accuracy of each user for general and agent-based user-adapted filtering mechanism. Note that the users are sorted in non-decreasing order of the number of tweets they chose to filter out of the 75 presented to them, i.e., higher numbered users chose to filter a higher percentage of the tweets they rated (the user number, on the x-axis, does not directly map into the percentage of tweets filtered).

We observe that for almost all the cases, user-adapted filter significantly outperform the general filter. The V-shape of the accuracy plot for the user-adapted filter reflects the fact that it is easier to predict the choice of users who chose to filter almost all or almost none of the tweets, compared to the user who were more discerning, e.g., chose to filter half of the tweets rated. We also find the accuracy rate of predicting the choice of the more discerning users to be more dispersed: choices of some of these users are significantly easier to predict than others. While the accuracy rates of filtering choices for the users in the middle range is not stellar, an online agent may be able improve its prediction with more interactions with the associated user. There are also a few odd instances where the general classifier do well; we believe those users may have filtering preferences more aligned with the "population average". As trends from the user-adapted and general filter trained using SVMs and Random Forest classifiers are similar, we have not included them.

6.3 Comparison of user-adapted classifier with majority filtering decision

We did further analysis with a Majority Class filter for each user which uses the majority filtering decision of a user, i.e., if a user chooses to filter a majority of the tweets presented then all tweets are filtered and vice versa. The purpose of using the Majority filter is to check whether an agent is able to leverage user-specific training data to build a user-adapted filtering mechanism and improve on a baseline filtering scheme which also is based only on that user's choices (as compared to the general filter in the previous section that is trained on data from all users). In Figure 7, the X-axis represents each user, ordered in increasing number of the tweets they chose to filter, and the Y-axis represents the accuracy of each user for a user-adapted Majority Class filter and the user-adapted filtering mechanisms. Figure 7 also shows the classification accuracy all three (NB, SVM and Random Forest) user-adapted learning classifiers. Table 3 which shows the number cases the majority classifier did better (win), worse, or tied with each of the learning filters. Results show that the user-adapted classifiers using Naive Bayes, Random Forest and SVM classifier does sligthly better than the user-adapted majority filter, whereas the Naive Bayes filter significantly outperforms it.

To further verify if the performance differences of the user-adapted learning and the majority filters are statistically significant, we use the Wilcoxon Signed Rank Test [38], which gives the following results: (a) the user-adapted filtering mechanisms based on Naive Bayes, Support Vector Machine and Random Forest performed significantly different from the user-adapted Majority class classifier, (b) the Naive Bayes filter performs significantly different than Random Forest and Support Vector Machine, (c) performance of Random Forest and Support Vector Machines are not statistically significantly different from each other.

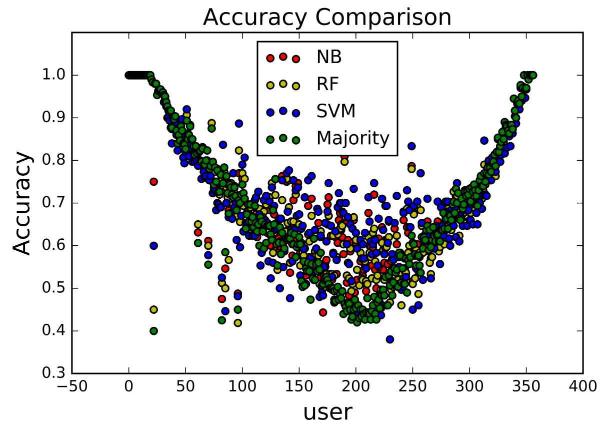


Figure 7: Accuracy of Majority class Vs (NB, SVM, RF).

Table 3: Comparison of performance of the learning algorithm with majority Algorithm.

Majority Vs	SVM	RF	NB
Win	149	132	52
Loose	166	143	136
tie	42	83	169

These results show that the user-adapted learning filtering mechanisms are able to leverage user-based training data to improve on an user-adapted baseline classifier.

7 RELATED WORK

Prior work on harassment detection spans several web platforms including Twitter, Facebook, Instagram, Yahoo!, YouTube, etc. Different platforms have unique communication modalities, user demographics and content and may, therefore, display different subtypes of hateful communication. For instance, one should expect quite different types of hate content on a platform catering to adolescents than on a web-platform used by a wider cross-section of the general public. Manual analysis of data and establishment of relationships between multiple features are often error-prone. Machine learning has been used to address this issue.

In [40], a supervised classification technique is used along with local, sentimental and contextual features extracted from a post using Term Frequency-Inverse Document Frequency (TF-IDF). The classification technique is conjugated with n-grams and other features, such as incorporating abusiveness, to train a model for detecting harassment. A significant improvement over the general TF-IDF scheme is observed while adding the sentimental and contextual features. In [32], Support vector machines (SVMs) were used to learn a model of profanity using the bag of words (BOW) approach to find the optimal features. This approach surpasses the performance of all previous list-based profanity detection techniques. A linguistic and behavioral pattern based model [24] was proposed to filter short texts, detect spam and abusive users in the network. It used real-world SMS data set from a large telecommunications operator from the US and a social media corpus. It also addressed different ways to deal with short text message challenges such as tokenization and entity detection by using text normalization and substring clustering techniques. A comprehensive approach to detecting hate speech was proposed in [35] which presents a plan that targets specific group characteristics, including ethnic origin, religion, gender, and sexual orientation. The paragraph2vec approach [13] is used to classify anti-Semitic speech on data collected over a 6-month period from Yahoo Finance website. In [26], a comprehensive lists of slurs, obtained from hate speech and an array of features for abusive language detection. such as POS tags, the presence of blacklisted words, n-gram features including token and character n-grams and length features, are used. Their scheme outperformed a deep learning approach by focusing on good annotation guidelines that help detect specific abusive language. In [8], an exploratory single blended model of cyber-hate that incorporates knowledge of features across multiple types was used. The proposed method improved classification for different types of cyber-hate beyond the use of a BOW and known hateful terms. In [36], author analyzed the impact of various extra-linguistic features in conjunction with character n-grams for hate speech detection. It was observed that though the gender feature is informative, differences in the geographic and word-length distribution do not improve performance over character level features. A list of criteria based on critical race theory to identify racist and sexist slurs was presented.

Most of the research studies that we have come across, including the ones summarized above, have focused on binary classification of communication. Cyber-harassment, however, is multi-faceted and may convey diverse content including hostility, humiliation, insults, threats, unwanted sexual advances, etc. to a target.

Human moderation efforts for curbing online abuse have proved to be inadequate and automated agents can be effective tools for mitigating cyber-harassment. Despite the research on automated harassment detection, quality research on automated intervention is sparse. Two recent studies, however, explore automated intervention: (i) [15] uses bots to help maintain block-lists, but requires manual user additions, and (ii) [1] uses a bot, imitating a person with similar demographic characteristics to the troll, to provide feedback to dissuade hateful comments. These approaches do not consider individual user's harassment perception or sensibility. In [30], author discussed some research challenges that can be approached with intelligent agents based solution. Author also sheds light on how an user-based agent can be useful to learn user's threat perception and acceptance and filter different abusive incoming communications directed towards the user.

As there is a gap between psychological theories and computational models and concepts of what constitutes aggression, training computers for comprehensive harassment detection has proved to be challenging. The goal of this research is to remove this disconnect and leverage computational and psychological approaches to identify different aggression categories, crowdsourced feedback, surveys and statistical techniques. Based on the identified categories, an effective agent-based detection tool has been built to filter communication based on user's sensibility and preferences.

8 CONCLUSIONS AND FUTURE WORK

Our goal is to develop a better understanding of cyber harassment types, the varying perception among the users to harassing communication, as well as the variation in the tolerance or acceptance of harassment categories, and hence the need for user-adapted filtering of harassing communication. To achieve our goal, we used a taxonomy of harassment categories grounded in the psychology literature, collected a set of tweets that matched identified keywords associated with different harassment categories, developed a crowd-sourced data-set where MTurk workers were asked to rate the harassment intensity of presented tweets and if they wanted to filtered them from their input stream. We performed an in-depth analysis of the labeled data to identify the variation of harassment sensibility and tolerance level among the users over different harassment categories. These results show how user's perspective varies depending on harassment categories and the intensity of the tweets. As different users' perceived intensity and acceptance could vary significantly for the same tweet, this highlighted the need for an adaptive learning agent that can user-adapted intervention mechanisms to mitigate the effects of cyber-harassment.

We trained agent-based user-adapted filters from labeled data and evaluated their ability to filter harassing communication for each user. Each agent learned to filter for a user using only about 60 tweets. The user-adapted filters performed better than one general filter trained on labeled data from all users. We also found that the learned user-adapted filters outperformed a base-line non-learning filter using the majority filtering choice for any tweet.

These results reaffirmed the pressing need for user-adapted harassment filtering mechanisms rather than relying on a single, generic filter for all. This work can be extended to provide supportive communication to the victims of cyber harassment/aggression. The victims of aggressive communication can be heartened by receiving supportive communication. A user based agent can identify the personality traits and the aggression tolerance levels and thus can provide supportive communication that dilute the chain of harassing comments received, divert attention from these negative inputs, and bolster the confidence and self-esteem of the target of aggression. A different and possibly effective future research avenue would be to develop an agent that sends responses to online aggressors with the goal of disarming, discouraging, and eliminating future harassment. Based on the learned user perception and tolerance, an agent can detect aggressive tweets and can identify the source. Based on other attributes like the frequency, intensity level, and the number of users attacked by such sources, an agent can detect potential threat. An agent can send commensurate responses to the aggressor or harasser to mitigate the situation and disarm the aggressor.

In the future, we can investigate unsupervised grouping of users of similar harassing preference and training a limited number, one per group, of user-adapted filters. An orthogonal research direction would be for adaptive agents to track any changing preference of the user and accordingly adapt the associated filter. This is particularly relevant as users change their attitude towards particular categories of harassment based on their personal life situation and interaction with new acquaintances.

REFERENCES

[1] 2016. Troll Hunters: The Twitterbots That Fight Against Online Abuse. http://cacm.acm.org/news/205680-troll-hunters-the-twitterbots-that-fight-against-online-abuse/fulltext. (2016). New Scientist, Communications of the ACM.

[2] Alan Agresti, Matilde Bini, Bruno Bertaccini, and Euijung Ryu. 2008. Simultaneous confidence intervals for comparing binomial parameters. *Biometrics* 64, 4 (2008), 1270–1275.

[3] Alana Barton and Hannah Storm. 2014. Violence and harassment against women in the news media: a global picture. *New York: WomenâĂŹs Media Foundation and the International News Safety Institute. Accessed November* 6 (2014), 2014.

[4] Anita Bernstein. 2014. Abuse and Harassment Diminish Free Speech. *Pace Law Review* 35, 1 (2014).

[5] Adam M Bossler, Thomas J Holt, and David C May. 2012. Predicting online harassment victimization among a juvenile population. *Youth & Society* 44, 4 (2012), 500–523.

[6] Danah Boyd, Scott Golder, and Gilad Lotan. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*. IEEE, 1–10.

[7] Erin E Buckels, Paul D Trapnell, and Delroy L Paulhus. 2014. Trolls just want to have fun. *Personality and individual Differences* 67 (2014), 97–102.

[8] Pete Burnap and Matthew L Williams. 2016. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science* 5, 1 (2016), 11.

[9] Marilyn A Campbell. 2005. Cyber Bullying: An Old Problem in a New Guise?. *Australian journal of Guidance and Counselling* 15, 01 (2005), 68–76.

[10] Pew Research Center. 2017. Social Media Usage: 2005-2017. http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2017/. (2017).

[11] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, 71–80.

[12] Kevin Coe, Kate Kenski, and Stephen A Rains. 2014. Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication* 64, 4 (2014), 658–679.

[13] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 29–30.

[14] Maeve Duggan. 2014. *Online harassment*. Pew Research Center.

[15] R Stuart Geiger. 2016. Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society* 19, 6 (2016), 787–803.

[16] Romulus Gidro, Aurelia Gidro, et al. 2016. Aspects Concerning Sexual And Moral Harassment In The Workplace. *Curentul Juridic, The Juridical Current, Le Courant Juridique* 64 (2016), 65–73.

[17] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications* 13, 4 (1998), 18–28.

[18] Sameer Hinduja and Justin W Patchin. 2010. Bullying, cyberbullying, and suicide. *Archives of suicide research* 14, 3 (2010), 206–221.

[19] Andreas M Kaplan and Michael Haenlein. 2010. Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons* 53, 1 (2010), 59–68.

[20] Chei Sian Lee and Long Ma. 2012. News sharing in social media: The effect of gratifications and prior experience. *Computers in Human Behavior* 28, 2 (2012), 331–339.

[21] Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomForest. *R news* 2, 3 (2002), 18–22.

[22] Andrew McCallum, Kamal Nigam, et al. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, Vol. 752. Citeseer, 41–48.

[23] Neil Middleton and Richard Schneeman. 2013. *Heroku: Up and Running: Effortless Application Deployment and Scaling*. " O'Reilly Media, Inc.".

[24] Alejandro Mosquera, Lamine Aouad, Slawomir Grzonkowski, and Dylan Morss. 2014. On Detecting Messaging Abuse in Short Text Messages using Linguistic and Behavioral patterns. *arXiv preprint arXiv:1408.3934* (2014).

[25] Elana Newman, Susan Drevo, Bradley Brummel, Gavin Rees, and Bruce Shapiro. 2016. Online abuse of women journalists: Towards an EvidenceâĂŘbased Approach to Prevention and Intervention. (2016), 46–52.

[26] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 145–153.

[27] Gwenn Schurgin O'Keeffe, Kathleen Clarke-Pearson, et al. 2011. The impact of social media on children, adolescents, and families. *Pediatrics* 127, 4 (2011), 800–804.

[28] Thomas E Page, Afroditi Pina, and Roger Giner-Sorolla. 2015. "It Was Only Harmless Banter!" The development and preliminary validation of the moral disengagement in sexual harassment scale. *Aggressive behavior* (2015).

[29] Jay P Paul, Joseph Catania, Lance Pollack, Judith Moskowitz, Jesse Canchola, Thomas Mills, Diane Binson, and Ron Stall. 2002. Suicide attempts among gay and bisexual men: lifetime prevalence and antecedents. *American journal of public health* 92, 8 (2002), 1338–1345.

[30] Sandip Sen, Zenefa Rahaman, Chad Crawford, and Osman Yücel. 2018. Agents for Social (Media) Change. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1198–1202.

[31] Lindsay H Shaw and Larry M Gant. 2002. In defense of the Internet: The relationship between Internet communication and depression, loneliness, self-esteem, and perceived social support. *Cyberpsychology & behavior* 5, 2 (2002), 157–171.

[32] Sara Owsley Sood, Judd Antin, and Elizabeth F Churchill. 2012. Using Crowdsourcing to Improve Profanity Detection.. In *AAAI Spring Symposium: Wisdom of the Crowd*.

[33] Frithjof Staude-Müller, Britta Hansen, and Melanie Voss. 2012. How stressful is online victimization? Effects of victim's personality and properties of the incident. *European Journal of Developmental Psychology* 9, 2 (2012), 260–274.

[34] Amazon Mechanical Turk. 2012. Amazon mechanical turk. *Retrieved August* 17 (2012), 2012.

[35] William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics, 19–26.

[36] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of NAACL-HLT*. 88–93.

[37] Nancy Willard. 2006. Cyberbullying and cyberthreats. *Eugene, OR: Center for Safe and Responsible Internet Use* (2006).

[38] RF Woolson. 2007. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials* (2007), 1–3.

[39] Michele L Ybarra. 2004. Linkages between depressive symptomatology and Internet harassment among young regular Internet users. *CyberPsychology & Behavior* 7, 2 (2004), 247–257.

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