

Forecasting the presence and intensity of hostility on Instagram using linguistic and social features

Ping Liu and Joshua Guberman

Illinois Institute of Technology
Chicago, IL 60616

Libby Hemphill

University of Michigan
Ann Arbor, MI 48109

Aron Culotta

Illinois Institute of Technology
Chicago, IL 60616

Abstract

Online antisocial behavior, such as cyberbullying, harassment, and trolling, is a widespread problem that threatens free discussion and has negative physical and mental health consequences for victims and communities. While prior work has proposed automated methods to identify hostile comments in online discussions, these methods work *retrospectively* on comments that have already been posted, making it difficult to intervene before an interaction escalates. In this paper we instead consider the problem of *forecasting* future hostilities in online discussions, which we decompose into two tasks: (1) given an initial sequence of non-hostile comments in a discussion, predict whether some future comment will contain hostility; and (2) given the first hostile comment in a discussion, predict whether this will lead to an escalation of hostility in subsequent comments. Thus, we aim to forecast both the presence and intensity of hostile comments based on linguistic and social features from earlier comments. To evaluate our approach, we introduce a corpus of over 30K annotated Instagram comments from over 1,100 posts. Our approach is able to predict the appearance of a hostile comment on an Instagram post ten or more hours in the future with an AUC of .82 (task 1), and can furthermore distinguish between high and low levels of future hostility with an AUC of .91 (task 2).

1 Introduction

Harassment, and the aggressive and antisocial content it entails, is a persistent problem for social media — 40% of users have experienced it (Duggan et al. 2014). Existing approaches for addressing toxic content, including automated mechanisms, crowdsourced moderation, and user controls, have so far been ineffective at reducing hostility. Addressing harassment and other hostile online behaviors is challenging in part because people (Guberman and Hemphill 2017), policies (Pater et al. 2016), and laws (Nocentini et al. 2010) disagree about what constitutes unacceptable behavior, it takes so many forms, occurs in so many contexts, and is started by many different kinds of users (Phillips and Milner 2017; Phillips 2015; Cheng et al. 2017).

Instead of trying to model all unacceptable content, we focus on a single class of content — *hostile comments* — that are common to a number of types of unwanted behavior including harassment and aggression. We define a hostile comment as one containing harassing, threatening, or offensive language directed toward a specific individual or group.

Automated detection using bots or machine learning to identify and possibly deter online harassment shows promise (Sood, Churchill, and Antin 2012; Geiger 2016; Reynolds, Kontostathis, and Edwards 2011), and would greatly reduce the burden on human moderators (both lay users and moderators employed by social media platforms). However, a key pitfall of existing automated methods is that they work *retrospectively* on comments that have already been posted, making it difficult to intervene before an interaction escalates. In other words, users and moderators must first be exposed to the content before it can be removed. Currently, no existing tools incorporate the affordances of predictive models to *prevent* harassment rather than *react* to it.

In this paper, we propose a method to forecast the arrival of hostile comments on Instagram posts. In order to support different intervention strategies, as well as to assess the difficulty of variants of this problem, we consider two forecasting tasks: (1) **hostility presence forecasting**: given the initial sequence of *non-hostile* comments in a post, predict whether some future comment will be hostile; (2) **hostility intensity forecasting**: given the first *hostile* comment in a post, predict whether the post will receive more than N hostile comments in the future. Solutions to the first task would support more aggressive interventions that attempt to eliminate all hostile comments from the system, while solutions to the second task would focus interventions on the most extreme cases.

We evaluate our approach on a newly constructed dataset of over 30k Instagram comments on over 1,100 posts. We identify a number of features computed over the current conversation as well as previous conversations that together provide a strong signal to forecast future hostilities in online communications. Our approach is able to predict the appearance of a hostile comment on a post ten or more hours in the future with an AUC of 0.82 (task 1), and can furthermore distinguish between high and low levels of hostility intensity with an AUC of 0.91 (task 2).

2 Related Work

In this section, we review prior methods for detecting hostile content online, as well as sociological literature motivating some of the predictive features used in this study.

2.1 Detecting hostile and antisocial content

The three primary approaches to detecting hostile content are (1) crowd-sourced reporting, (2) moderators, (3) automated detection algorithms. Most websites, blogs, and social media platforms crowdsource reporting and rely on their users to report or flag content as being abusive or in violation of the platform’s terms and conditions. Many systems also include blocking controls that allow users to prevent others from viewing or commenting on their content. Human commercial content moderators may then attend to the flagged posts (Crawford and Gillespie 2014). Crowdsourced bot-based Twitter blocklists (Geiger 2016) combine automation, crowdsourcing, and user controls to help users moderate their feeds.

The flagging or reporting approach to responding to harassment exposes not only users to potentially abusive content but also the moderators who respond to it (Roberts 2016). Also, like any collaborative filtering mechanism, crowdsourcing harassment detection runs the risk of shilling attacks (Su and Khoshgoftaar 2009) where people give positive recommendations to their own materials and negative recommendations to others’. These reporting mechanisms also rely on free labor from members of marginalized communities, and often result in the reporters being targeted for additional harassment (Nakamura 2015).

Automated algorithms may also be used to augment or replace crowdsourced moderation. In theory, automated classification of posted content would permit social media platforms to respond to greater quantities of reports in less time, thus minimizing users’ exposure. To this end, effective classifiers would allow platforms to keep abreast of toxic content. Automated mechanisms are the least common approach to addressing hostile content currently, but Instagram recently introduced tools to fight spam and encourage kinder commenting (Thompson 2017).

Prior work in this area has focused on the classification of existing, rather than future, content (Yin et al. 2009). A number of classification systems have been proposed to identify toxic content, using standard text features such as TF-IDF, lexicons, and sentiment, as well as social-network features, such as number of friends/followers, and demographic attributes (Yin et al. 2009; Dinakar, Reichart, and Lieberman 2011; Reynolds, Kontostathis, and Edwards 2011; Sood, Churchill, and Antin 2012; Dadvar et al. 2012; Bellmore et al. 2015; Geiger 2016; Al-garadi, Varathan, and Ravana 2016).

In this paper, we consider the more ambitious problem of forecasting future hostile content, which has the potential to allow stakeholders to stay *ahead* of some toxic content. Doing so would allow alternative intervention strategies, such as issuing alerts to indicate an immediate threat and to trigger both social and technical responses to potential hostilities. Task 1’s results are especially useful here where we are

able to forecast whether non-hostile comments will be followed by hostile comments. Task 2’s results can additionally help prioritize interventions to posts that are likely to see a sharp rise in hostile comments.

Few attempts at forecasting have been made so far, but some researchers have developed features that may be useful for this task. For instance, Wang and Cardie (2016) classify *disputes* using lexical, topic, discussion, and sentiment features. One of their features, *sentiment transition*, estimates the probability of sentiment changing from one sentence to the next. The feature is used to augment the classifier’s overall labeling capabilities. That is, rather than attempting to use sentiment transition probability to predict whether there will be a dispute within the next N comments, the feature is used to bolster the classification of the sentiment within existing content.

Recently, Cheng et al. (2017) examined whether features of the context of a conversation and/or users involved could predict whether a future post will be flagged for removal by a moderator in a news discussion board. They found that an individual’s mood and contextual features of a given discussion (e.g., how recently others had posted flag-worthy comments, how much time had passed since their own last comment) are related to an individual’s likelihood of writing a flagged comment. Relating to discussion context, they found that an initial flagged comment predicts future flagged comments, and that the likelihood of a new discussant posting a trolling comment increases with the number of flag-worthy comments already present within a thread. While similarly motivated, there are several important distinctions between this past work and the present study: first, we forecast hostility, rather than flagged posts; second, the best model in Cheng et al. (2017) assumes knowledge of the username of the author of the future flagged post, which, while useful for understanding the nature of trolling, is not a feature that would be observable by a real forecasting system. (It requires knowing who will comment on a post in the future.)

2.2 Sociological indicators of conflict

The literature on conflict and escalation reveal that certain patterns of interaction are more likely to lead to hostile communication, and we therefore include features that capture patterns in the comment thread. For instance, a “conflict spiral” (Rubin, Pruitt, and Kim 1994) occurs when parties mirror each other’s aggressive communication tactics meaning that no matter who posts the first aggressive comment, other users will likely follow with aggressive comments of their own. Research on empathy in online communities suggests that people rarely ask for help directly (Preece 1998; Pfeil and Zaphiris 2007), but it is likely that help begets help in the same way that aggression begets aggression. In our Instagram scenario, this means that conversations where hostile comments are followed by positive comments may be able to recover rather than devolve into more hostility.

Conflict spirals are less likely to occur when social bonds between parties can mediate the escalation (Rubin, Pruitt, and Kim 1994). Social similarity also mediates conflict — in part because it is easier for us to empathize with people who are similar (Krebs 1975). Relatively homogeneous groups

| | posts | comments | hostile comments |
|--------------------------|-------|----------|------------------|
| hostile posts | 591 | 21,608 | 4,083 |
| non-hostile posts | 543 | 9,379 | 0 |
| total | 1,134 | 30,987 | 4,083 |

Table 1: Statistics of the Instagram posts. Hostile posts are those with at least one hostile comment.

in which people see themselves as “in-group members” are less likely to experience hostility as well (Dovidio et al. 2013). Together, social similarity and in-group preferences mean that posts with homogeneous commenters should be less likely to receive many hostile comments. Our conversation trend and user similarity features (c.f., §5) enable us to examine the predictive utility patterns of hostility and social similarity.

Prior behavior is a useful indicator of future behavior. For instance, as mentioned above, Cheng and colleagues found that a user’s previous posting history was a reasonable predictor of their future posts but that the context in which they post were better predictors (Cheng et al. 2017). Similarly, Chen et al. (2012), classify both users and content, recognizing that features of both are potentially useful for reducing antisocial behaviors. We include measures of both authors’ and commenters’ previous posts and use different measures of time and comment thread patterns.

Many toxic remarks contain various types of profanity (Märtens et al. 2015; Wang et al. 2014) or hate speech (Davidson et al. 2017); we use existing lexicons to generate features for both. However, we recognize that what constitutes hate speech and profanity depends on the context and geography of use (Sood, Antin, and Churchill 2012; Davidson et al. 2017).

In summary, the vast majority of prior work has focused on identifying hostile messages after they have been posted; here, we focus on forecasting the presence and intensity of future hostile comments. The primary contributions of this paper, then, are as follows:

- The creation of a new corpus of over 30K Instagram comments annotated by the presence of hostility (§3).
- A temporal clustering analysis to identify distinct patterns of hostile conversations (§4).
- The formulation of two new tasks and machine learning solutions to forecast the presence and intensity of hostile comments on Instagram posts (§5).
- A thorough empirical analysis to understand the strengths and weaknesses of the forecasting algorithms (§6-7).

3 Data collection and annotation

Given the prevalence of online harassment among teenagers (Duggan et al. 2014; Duggan 2017), as well as the vulnerability of this population, we decided to target our data collection to this group. According to local

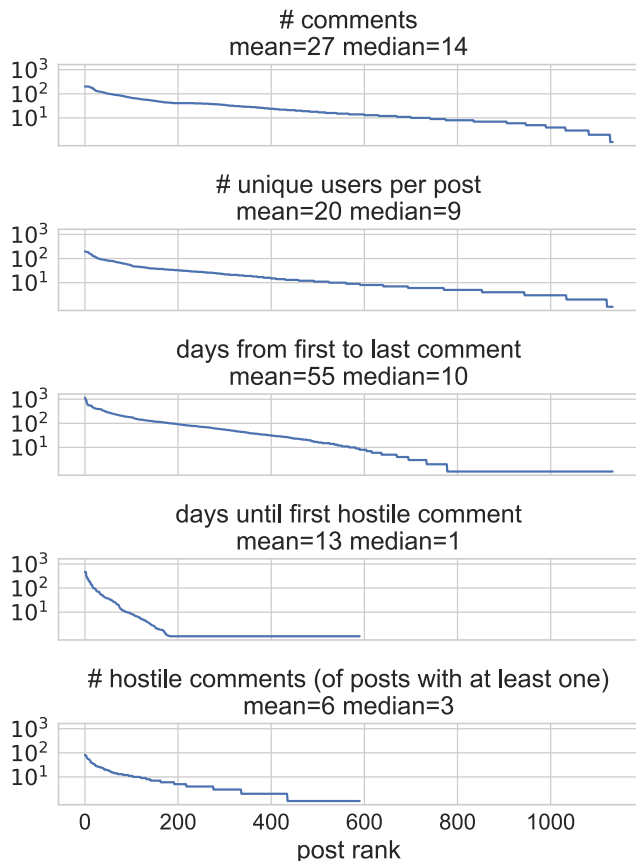


Figure 1: Rank-count plots of statistics of posts.

schools, parent groups, and recent surveys (NORC 2017), Instagram is one of the most popular social media sites among teenagers. Instagram’s predominantly public content and associated comments make it possible to collect and annotate the content necessary for our experiments.¹

We refer to a user-provided image and its associated comments as a *post*. Each post is authored by one user, and each post may have hundreds of comments from other users. Each comment carries a username and timestamp. To collect a sample of posts from this population, we first seeded a web crawler with the usernames of users who have public accounts, and who, upon manual inspection, appear to match the target age demographic, live in the United States, and post predominately in English. We then collected their most recent posts and associated comments. We then conducted snowball sampling by adding usernames discovered in each comment to a queue of additional users to crawl. This resulted in over 15M comments.

For annotation purposes, we defined a hostile comment as one containing harassing, threatening, or offensive language directed toward a specific individual or group. We focus on this type of content as it is a common characteristic in many different types of antisocial, unwanted internet behavior and

¹All data and research protocols in this paper were approved by the relevant Institutional Review Board.

is a high priority for intervention. For training and validation, we required a sample of posts for which all comments are labeled with a binary indicator of hostility.

As with many other classification problems with high class imbalance, uniform sampling would result in a very low fraction of positive instances. To more efficiently allocate annotation resources, we first built a search interface to interactively explore the 15M comments retrieved via snowball sampling. After several rounds of searching and browsing, we identified a broad selection of keywords that frequently occur in hostile comments (e.g., certain profanities, abbreviations, and emojis). We then uniformly sampled 538 comments matching at least one of the identified keywords, and then downloaded all the comments from the parent post of each comment for annotation, resulting in over 20K comments. Additionally, because the experiments also required posts without any hostile comments, we sampled for annotation an additional 596 posts that did not contain any hostility keywords. (Due to the nature of our crawling procedure, we draw these non-hostile posts from similar user populations as the hostile posts.) Altogether, this resulted in 1,134 posts with 30,987 comments for annotation.

These 30,987 comments were annotated by Amazon Mechanical Turk workers who had passed certification tests indicating agreement with our labeling of example content. Each comment was annotated by two independent workers, with disagreements resolved by a third worker when needed.

For the annotation tasks, workers were presented with an Instagram post (consisting of an image and, in some cases, a caption) and comments written in response to the post. We instructed workers to determine whether a comment was innocuous, hostile or offensive, or contained a physical threat. We also asked workers to mark comments that appear to be *flashpoints*—points within already-hostile conversations at which hostility escalates precipitously, or at which a non-hostile conversation becomes markedly hostile. We directed workers to make judgments based on users’ presumed intent to cause harm or distress to other users. Overall, workers were in agreement 87% of the time when assessing whether a comment was hostile (Cohen’s $\kappa = .5$).

Roughly half (591) of the posts contain at least one hostile comment, and the remainder (543) have no hostile comments. Across all data, 4,083 of 30,987 comments (13%) are hostile. (See Table 1.)

Figure 1 shows additional data statistics as rank-count plots (which rank posts in descending order of each statistic).

All five statistics — comments per post, hostile comments per post, time-duration of conversations on a post, time until the first hostile comment, and unique users per post — exhibit long-tail properties. We also note that discussions on a post extend over long time periods — the median post receives its final comment 10 days after the post was created. Similarly, while the first hostile comment often appears the day the post is created, 20% of hostile posts do not receive a hostile comment until 5 or more days after they were created.

In summary, these data suggest that the commenting feature on Instagram is heavily used, with posts receiving many

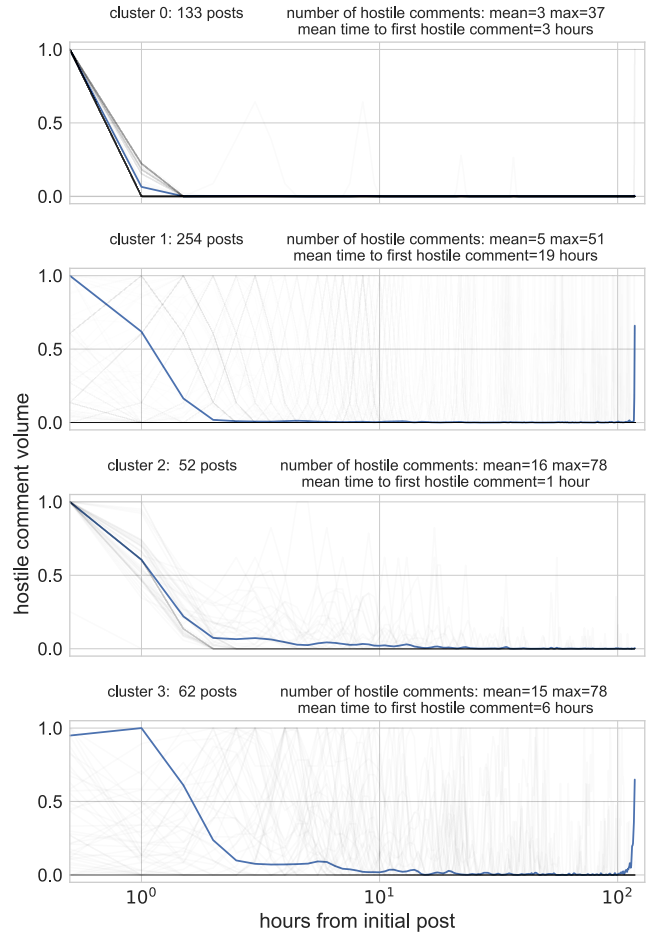


Figure 2: Volume of hostile comments over time for four discovered clusters of posts.

comments from many users over long periods of time. Furthermore, among posts containing at least one hostile comment, the volume and temporal frequency of hostile comments appear to vary substantially between posts, which motivates our tasks of forecasting both the presence and intensity of hostile comments.

4 Identifying hostility patterns suitable for forecasting

To motivate the forecasting tasks, we first explored variations in the onset and volume of hostile comments to determine which sorts of posts may be amenable to forecasting. To do so, we performed a cluster analysis to identify groups of posts that exhibit similar temporal characteristics. We first constructed a time series for each post by counting the number of hostile comments posted within each hour from the time the post was created.² We restricted this analysis to the 501 posts that have at least one hostile comment within the first 10 days. We then clustered the time series using the K-Spectral Centroid (K-SC) clustering algorithm of Yang and

²We smooth each series with a Gaussian kernel (width=5, $\sigma=1$).

Leskovec (2011). K-SC clusters a set of time series using a distance measure that is invariant to scaling and shifting, enabling us to identify posts that have similarly shaped time series of hostile post frequency.

Figure 2 shows four discovered clusters of posts.³

Each gray line corresponds to the smoothed time series for a single post, and the blue lines indicate the centroid of each cluster. The y-axis is scaled to 1 by dividing each series by its maximum value. We can see four distinct patterns. Cluster 0 has the lowest level of hostility (only three hostile comments, on average), and these hostile comments arrive almost immediately after the post is created. Cluster 1 also has a relatively small number of hostile comments, but these can arrive much later (19 hours after the post was created, on average). Clusters 2 and 3 show a similar pattern, but for posts with roughly three times the volume of hostile posts. Thus, clusters 1 and 3 both have a delay in the onset of the first hostile message, but cluster 3 results in a much higher volume of hostile messages in the remainder of the comments.

These results provide some insight into what sorts of interventions may have an impact and what their limitations may be. For example, posts from cluster 0 appear to be a low priority for intervention — the hostile comments are isolated incidents that occur too quickly for any preventative action. In contrast, in clusters 1 and 3 the time lag before hostile comments appear presents an opportunity to intervene before hostilities escalate — e.g., by blocking comments, taking down a post, or other means. This motivates our first task of forecasting a future hostile comment in a post. Additionally, while the first hostile comment appears similarly quickly in clusters 0 and 2, if we could distinguish between the two after the first hostile comment appears, we could prioritize posts from cluster 2 for intervention. This motivates our second task of forecasting the total volume of hostile comments a post will eventually receive.

5 Forecasting hostility

In this section, we formalize our two forecasting tasks and our proposed approach. Our data consist of m posts $\{p_1 \dots p_m\}$, where post p_i is an ordered list of tuples $p_i = \{(c_1, t_1, y_1) \dots (c_n, t_n, y_n)\}$, with (c_j, t_j, y_j) indicating the j th comment, timestamp, and hostility label, respectively. We assume $y_j = 1$ indicates a hostile comment, and $y_j = 0$ otherwise. Given the patterns observed above, we consider the following two forecasting tasks:

Task 1: Hostility presence forecasting. Given an initial sequence of *non-hostile* comments on a post, predict whether at least one future comment will be hostile. That is, for each post the system observes $\{(c_1, t_1, y_1) \dots (c_j, t_j, y_j)\}$ where $y_i = 0 \forall i \leq j$ and must predict whether there is some subsequent comment $k > j$ such that $y_k = 1$.

We refer to $t_k - t_j$ as the *lead time*: the amount of time between the final comment observed by the system and the time of the first hostile message. From an intervention perspective, greater lead times are desirable, since they provide

more opportunity to take preventative measures. However, we expect problem difficulty to increase with required lead times, which we investigate below.

Task 2: Hostility intensity forecasting. Given all comments up to and including the first hostile comment, predict whether the total number of hostile comments on a post will be greater than or equal to a given threshold N . That is, the system observes $\{(c_1, t_1, y_1) \dots (c_j, t_j, y_j)\}$, where $y_j = 1$ and $y_i = 0 \forall i < j$. A post containing m comments has a positive label if $\sum_{k=1}^m y_k \geq N$. When N is large, the system is designed to identify posts with a high intensity (volume) of hostile comments, which can then be prioritized for intervention.

Because we frame both tasks as binary classification problems, we model each task using logistic regression with L2 regularization.⁴ We next describe the features used for this classifier.

5.1 Features

For both tasks, we compute features over the linguistic context of the current post and the historical context of the users participating in the post.

- **Unigram (U)** We tokenize each comment into unigrams, retaining emojis as separate words, and collapsing all user @mentions to a single feature type. Multiple comments are collapsed into a single bag-of-words vector.
- **Word2vec (w2v)** To overcome the sparsity of unigram features, we fit a word2vec model (Mikolov et al. 2013) on the ~ 15 M unlabeled comments collected in our initial crawl. We set the vector length to 100 dimensions, and to aggregate the vectors for each word in a comment, we concatenate the maximum and average along each word vector dimension (i.e., 200 total features).
- **n-gram word2vec (n-w2v)** Bojanowski et al. (2016) introduce an extension to word2vec designed to model character n-grams, rather than words, which may help when dealing with short social media messages that contain abbreviations and spelling errors. We train on the same 15M comments using this n-gram word2vec model and again use maximum and average vectors to aggregate across words in a sentence.
- **Hatebase/ProfaneLexicon (lex)** We include features indicating whether a comment contains words present in two lexicons of offensive language: Hatebase⁵ and Luis von Ahn’s Offensive/Profane Word List⁶. Hatebase has 582 hate words in English, and categorizes each hate word into six different categories: class, disability, ethnicity, gender, nationality and religion. For each category, we have a binary feature that is one if the sentence has at least one word in Hatebase; we add an additional feature for the total count across all categories. The Offensive/Profane

⁴Additional experiments with a Random Forest classifier produced similar results.

⁵www.hatebase.org

⁶www.cs.cmu.edu/~biglou/resources/

³Increasing the number of clusters up to 10 had similar results.

Word List provides 1,300+ English terms that may be considered offensive; we include both binary and count features for this lexicon as well.

- **Final comment (final-com)** For posts with many comments, it may be necessary to place greater emphasis on the most recent comment. To do so, we create a separate feature vector using only the most recent comment, using unigram, n-gram word2vec, and lexicon features. E.g., one such feature may indicate that a word from the Hatebase lexicon appeared in the most recent observed comment for this post.
- **Previous comments (prev-com)** Prior work suggests that a user’s recent posts may indicate their propensity to post offensive comments in the near future (Cheng et al. 2017). To capture this notion, we first identify the usernames of the authors of each comment the system observes in the target post. We then identify the most recent comment each user has made in a post other than the target post. We then concatenate all these comments together, building unigram, n-gram word2vec and lexicon features. These features serve to summarize what the commenters in the current post have recently said in prior posts.
- **Previous post (prev-post)** Conversely, certain users’ posts may be more likely to receive hostile comments than others, perhaps because of their content, recent adversarial interactions with other users, or other attributes of the user. To capture this, we collect the most recent post made by the author of the target post and concatenate all the comments in that post, again using unigram, n-gram word2vec and lexicon features. These features serve to summarize the recent conversations involving the author of the target post.
- **Trend Feature (trend)** While the linguistic features summarize the context of the current conversation, they ignore the order of the comments. The goal of this feature type is to detect an increase over time in the hostilities expressed in the current post to identify trends that precede hostile comments. To do so, we first train a separate logistic regression classifier to determine if a comment expresses hostility or not, using the same unigram, n-gram word2vec, and lexicon features from above (it obtains cross-validation AUC of ~ 0.8). We then apply this classifier to each comment observed in the target post, recording the predicted posterior probability that each comment is hostile.⁷ We then compute several features over this vector of posterior probabilities: (1) the number and fraction of comments that have positive probability larger than some threshold (0.3 in the experiments below); (2) the maximum slope within two adjacent probabilities of positive class (to detect sudden shifts in the conversation); (3) the difference between the maximum and minimum probability across all comments.
- **User activity (user)** Finally, inspired by evidence that group heterogeneity can increase likelihood of hostility (Dovidio et al. 2013), we compute two additional features

⁷This classifier is trained with double cross-validation so that it never predicts the label of a comment that was in its training set.

based on the number of users who have commented on the target post: (1) the number of unique users divided by the number of comments (to quantify the heterogeneity of comment authors); and (2) the fraction of comments containing a user @mention (to quantify the volume of directed utterances).

6 Experiments

We conduct 10-fold cross-validation experiments to measure the forecasting accuracy for each task.⁸ For Task 1, the system observes the first comments of a post and predicts whether some hostile comment will appear. Rather than fixing the number of comments observed, we instead vary the lead time, defined as the time from the final comment observed on a post to the time of the first hostile comment. For example, if the first hostile comment on a post occurs at 3 p.m., a lead time of 1 hour means that the forecasting system will observe all comments on the post made before 2 p.m. We consider lead times of $\{1, 3, 5, 8, 10\}$ hours. Since the optimal classifier may vary by lead time, we train separate models for each lead time. We chose these lead time intervals based on our observations of the pace at which comments are posted and when hostile comments first appear (see Figure 2).

For some lead times, posts may be discarded from experiments — e.g., if the lead time is 10 hours, but the first hostile message occurs in the first hour, then it is not possible to observe any comments in time for that post. Similarly, the lead time also affects the number of comments the system may observe on average (i.e., longer lead times typically mean fewer comments may be observed before making a prediction). We sample an equal number of positive and negative examples for this task; to avoid introducing bias between classes, positive and negative samples are matched by the number of comments (e.g., for a positive example where we observe six comments, we sample a matched negative example from the set of non-hostile posts and restrict the observation to the first six comments.)

For the second task (hostility intensity forecasting), the system observes the first comments of a post, up to and including the first hostile comment, and predicts whether there will be $\geq N$ total hostile comments on the post. Thus, N serves as a threshold for instances to be considered in the positive class. From a practical perspective, it is more useful to distinguish between posts with 1 versus 10 hostile comments rather than 1 versus 2 hostile comments. We thus vary N from 5-15, training separate models for each setting. For a given N , positive instances must have $\geq N$ hostile comments, while negative instances must have only 1 hostile comment (posts with other values are not included in this analysis). Thus, the goal is to separate “one-off” hostile comments from those that result in an escalation of hostility.

Because both tasks can be viewed as ranking problems (to prioritize interventions), our primary evaluation metric is AUC (area under the ROC curve). For comparison, we also report F1, Precision, and Recall.

⁸For replication code, see <https://github.com/tapilab/icwsm-2018-hostility>

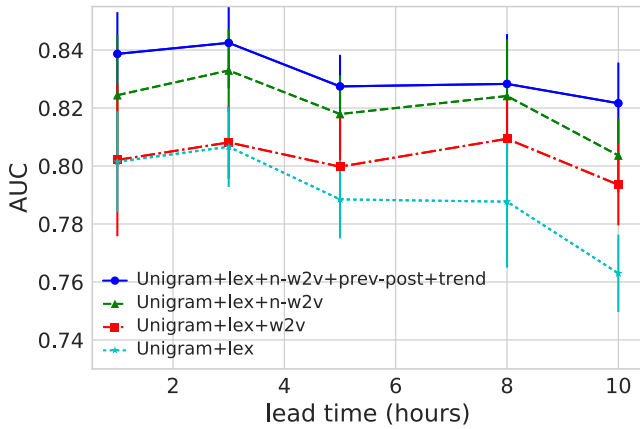


Figure 3: Hostility presence forecasting accuracy as lead time increases.

| Features | AUC | F1 | Precision | Recall |
|----------------------|--------------|--------------|--------------|--------------|
| Unigram | 0.790 | 0.737 | 0.723 | 0.778 |
| U + prev-com | 0.707 | 0.672 | 0.660 | 0.693 |
| U + final-com | 0.779 | 0.729 | 0.716 | 0.757 |
| U + trend | 0.790 | 0.739 | 0.725 | 0.776 |
| U + user | 0.791 | 0.742 | 0.729 | 0.778 |
| U + lex | 0.792 | 0.750 | 0.732 | 0.788 |
| U + w2v | 0.794 | 0.725 | 0.715 | 0.749 |
| U + n-w2v | 0.810 | 0.736 | 0.725 | 0.746 |
| U + prev-post | 0.828 | 0.761 | 0.756 | 0.765 |
| Best | 0.843 | 0.765 | 0.755 | 0.778 |

Table 2: Forecasting accuracy of Task 1 (lead time = 3 hours). The best combination uses all features except for w2v and prev-com.

7 Results

7.1 Task 1: Presence forecasting

Figure 3 shows the AUC for four feature settings as the lead time increases. The best model achieves a high AUC for a lead time of one hour (0.84). There is a steady drop in accuracy as lead time increases, as expected. However, even with a 10 hour lead time, the best model only declines by 2% from a lead time of one hour (0.84 to 0.82). This suggests that there is sufficient linguistic context early in a post to indicate whether it will eventually attract hostile comments. This figure also suggests that the n-gram word2vec features are somewhat more predictive than the traditional word2vec features across all lead times.

Table 2 provides a detailed feature comparison for a lead time of 3 hours. We can see that the previous post features are the most useful addition to the unigram features, increasing AUC from .790 to .828. Thus, users who have received hostile comments in the past are more likely to receive them in the future. In contrast, the previous comment features are unhelpful — including them actually reduces AUC from .790 to .707. Upon further inspection, we attribute this to

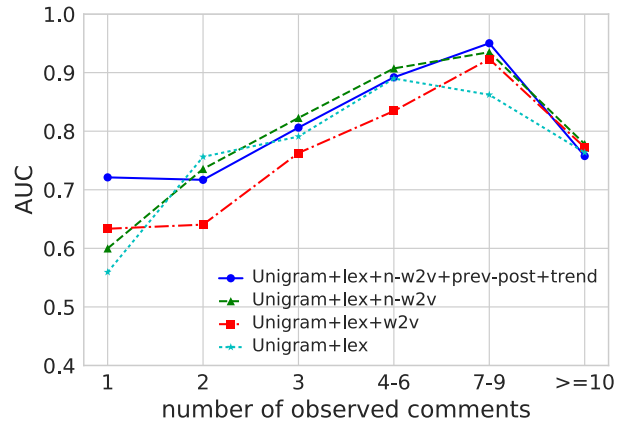


Figure 4: Hostility presence forecasting accuracy as observed comments increase (lead time=3 hours).

the fact that very often (494 of 591 hostile posts) the first hostile comment on a post is from a user who has not previously commented on this post. Thus, including features from the users who have commented on the post already may tell us little about the user who will initiate hostility. Together, these results suggest that certain users are more susceptible to receiving hostile comments.

Figure 4 stratifies the results of the four models by the number of observed comments, where the lead time is 3 hours. We can see that the number of observed comments can greatly affect accuracy. The best model has an AUC of .71 when observing only one comment, but an AUC of .92 when observing 7-9 comments. In practice, one may wish to restrict forecasts to posts with a sufficient number of observed comments. Furthermore, it appears that the previous post features most improve over baselines when there are few observed comments, presumably by providing additional context that is missing when there are few comments.

Figure 4 somewhat surprisingly shows a decrease in AUC when observing 10 or more comments. By reviewing these posts, we observe several instances of what we call “non-sequitur hostility.” For example, a post may quickly receive 20 innocuous, non-controversial comments (e.g., congratulating the poster on some achievement). Subsequently, a new user writes a hostile comment that is unrelated to the preceding conversation. These instances are very difficult to predict — since the model observes many innocuous comments, it assigns a low probability of seeing a hostile comment in the future. It is possible that accuracy may be improved by training separate classifiers after grouping by the number of observed comments.

7.2 Task 2: Intensity forecasting

Figure 5 shows the AUC for the best model as the positive class threshold increases. As expected, it is easier to identify posts that will receive 10 or more hostile comments than those that will receive only 5 or more. Table 3 reports results from different feature combinations using 10 as the positive class threshold. As in Task 1, the previous post features lead to improved AUC (.842 vs .808), but the previous comment

| Features | AUC | F1 | Precision | Recall |
|----------------------|--------------|--------------|--------------|--------------|
| Unigram | 0.808 | 0.747 | 0.741 | 0.673 |
| U + w2v | 0.753 | 0.696 | 0.662 | 0.673 |
| U + prev-com | 0.786 | 0.701 | 0.694 | 0.605 |
| U + user | 0.817 | 0.761 | 0.752 | 0.695 |
| U + n-w2v | 0.821 | 0.775 | 0.781 | 0.711 |
| U + trend | 0.825 | 0.778 | 0.782 | 0.721 |
| U + lex | 0.827 | 0.776 | 0.785 | 0.705 |
| U + prev-post | 0.842 | 0.782 | 0.829 | 0.688 |
| U + final-com | 0.879 | 0.792 | 0.805 | 0.722 |
| Best | 0.913 | 0.805 | 0.785 | 0.772 |

Table 3: Forecasting accuracy of Task 2 (N=10). The best feature combination is trend/user/final-com.

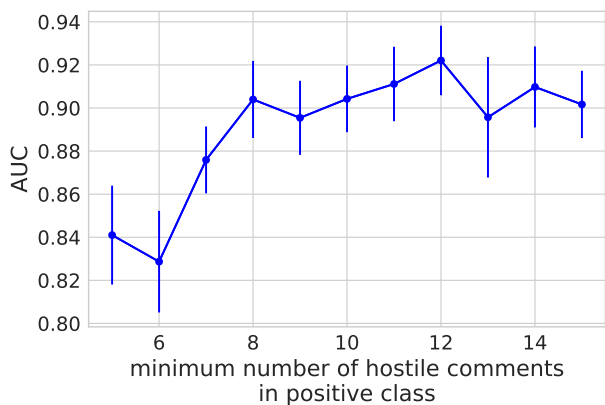


Figure 5: Hostility intensity forecasting accuracy as the positive class threshold increases (with standard errors).

features do not (.786 vs. .808). These results suggest that it is possible to distinguish between posts that will receive an isolated hostile comment and those that will receive high volumes of hostile comments. Thus, interventions may be prioritized by the expected intensity of hostile interactions.

To better understand the which features are leading indicators of a high volume of hostile comments, we examined the top coefficients for the positive class for $N = 10$. Overall, the final-comment features are the most predictive — it seems that features of the first hostile message are predictive of the volume of hostile message in the future. Examining the top coefficients for these features, we find that if the first hostile comment contains a user mention, a second-person pronoun, or profanity, it is correlated with a high intensity post. Other top features include: (1) a term matching the Profane Lexicon was used in a comment made on an earlier post by this author; (2) a comment on the current post contains a term from the Gender category in the HateBase lexicon; (3) the abbreviation “stfu” (“shut the f**k up”), which indicates a possible turning point in the conversation leading to an escalation in hostilities; (4) singular second- and third-person pronouns like “you” and “she,” which indicate that the con-

versation is directed toward specific users rather than distant third parties.

We also examined a sample of forecasting errors. False negatives often occur when many similar innocuous messages are posted consecutively, giving no indication of any escalation. For example, in one post, the first thirty comments were short messages wishing the author a happy birthday. The first hostile comment insulted the post author, resulting in many retaliatory messages from his friends. However, given the limited information in the first comments, the system did not predict this escalation in hostility.

This example demonstrates the nuance of online hostilities — it is rare in our data to see clear victims and perpetrators. Instead, it is more common to see an outsider write insulting messages to an individual, which in turn leads others to make insulting responses, which in turn escalates into a back-and-forth of hostile messages.

8 Limitations

As described in §3, the data is not a uniform random sample of all Instagram posts. Thus, the observed frequency of hostile comments should not be interpreted as an estimate of the true fraction of hostile comments on Instagram. Instead, we focused our limited annotation resources on posts that are (1) likely to be from vulnerable populations (teenagers) and (2) likely to contain hostility. To do so, we used a broad set of keywords to find an initial set of 538 comments (and corresponding posts), which we then augmented with 596 posts that did not contain such keywords. Although the seed sample of 538 comments is influenced by the keywords chosen, by annotating all the other comments in each post we also capture many additional hostile comments that do not match the initial set of keywords. Indeed, over 85% of the hostile comments in the final data were not retrieved by the initial keyword search. Furthermore, of the 538 posts containing a hostile keyword, only 504 were annotated as hostile; of the 596 posts that did not contain a hostile keyword, 87 had at least one comment annotated as hostile. Thus, the data contain many hostile comments that do not match the keywords, as well as many non-hostile comments that do match the keywords.

While the data may not be strictly representative of the population of all types of hostile content, we do find it to reflect a broad array of hostile content among a demographic that is highly vulnerable to the effects of online hostility.

9 Conclusion

We proposed methods to forecast both the presence and intensity of hostility in Instagram comments. Using a combination of linguistic and social features, the best model produces an AUC of 0.82 for forecasting the presence of hostility ten or more hours in the future, and an AUC of 0.91 for forecasting whether a post will receive more than 10 hostile comments or only one hostile comment. We find several predictors of future hostility, including (1) the post’s author has received hostile comments in the past; (2) the use of user-directed profanity; (3) the number of distinct users participating in a conversation; and (4) trends in hostility thus far

in the conversation.

By distinguishing between posts that will receive many hostile comments and those that will receive few or none, the methods proposed here provide new ways to prioritize specific posts for intervention. Since moderation resources are limited, it makes sense to assign them to posts where there is still time to de-escalate a conversation or prevent additional commenting. For instance, Instagram and similar platforms could use our approach to manage their moderation queues. Similarly, Instagram may use features our approach identifies (e.g., prior hostilities, number of distinct users) to improve their existing comment management features—offensive comment filtering and comment controls.

There are several avenues for future work. Given that Instagram is primarily a photo-sharing site, it is possible that image classification algorithms can be used to identify image attributes that predict hostile comments. Second, given the predictive power of the previous post feature, inferring more detailed user attributes (e.g., age, gender, ethnicity, etc.) may provide additional context for forecasting. Third, it may be possible to extract from these data insights into the effectiveness of de-escalation strategies — e.g., given a hostile comment, are there certain responses that diffuse the situation, as opposed to leading to increased hostilities? Argumentation mining methods may be applicable here (Palau and Moens 2009). Finally, we are experimenting with user-facing tools that allow users to monitor their own feeds (or their children's) for brewing escalations, enabling them to take action on- or offline.

10 Acknowledgements

This material is based upon work supported by the Nayar Prize at the Illinois Institute of Technology and the National Science Foundation under Grant No. 1822228.

References

- Al-garadi, M. A.; Varathan, K. D.; and Ravana, S. D. 2016. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network. *Computers in Human Behavior* 63:433–443.
- Bellmore, A.; Calvin, A. J.; Xu, J.-M.; and Zhu, X. 2015. The five ws of bullying on twitter: who, what, why, where, and when. *Computers in human behavior* 44:305–314.
- Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.
- Chen, Y.; Zhou, Y.; Zhu, S.; and Xu, H. 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 Conference on Social Computing (SocialCom)*, 71–80. IEEE.
- Cheng, J.; Bernstein, M.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. 1217–1230. ACM Press.
- Crawford, K., and Gillespie, T. 2014. What is a flag for? social media reporting tools and the vocabulary of complaint. *New Media & Society* 1461444814543163.
- Dadvar, M.; de Jong, F.; Ordelman, R.; and Trieschnigg, R. 2012. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*, 23–26. University of Ghent.
- Davidson, T.; Warmley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language.
- Dinakar, K.; Reichart, R.; and Lieberman, H. 2011. Modeling the detection of textual cyberbullying. In *Fifth International AAAI Conference on Weblogs and SocialMedia*.
- Dovidio, J. F.; Hewstone, M.; Glick, P.; and Esses, V. M., eds. 2013. *The SAGE Handbook of Prejudice, Stereotyping and Discrimination*. Los Angeles: SAGE Publications Ltd, 1 edition edition.
- Duggan, M.; Rainie, L.; Smith, A.; Funk, C.; Lenhart, A.; and Madden, M. 2014. Online harassment. washington, dc: Pew research center.
- Duggan, M. 2017. Online harassment 2017. <http://www.pewinternet.org/2017/07/11/online-harassment-2017/>. Accessed: 2018-1-26.
- Geiger, R. S. 2016. Bot-based collective blocklists in twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication, and Society* 19(6):787–803.
- Guberman, J., and Hemphill, L. 2017. Challenges in modifying existing scales for detecting harassment in individual tweets. In *50th Annual Hawaii International Conference on System Sciences (HICSS-50)*.
- Krebs, D. 1975. Empathy and altruism. *J. Pers. Soc. Psychol.* 32(6):1134–1146.
- Märtens, M.; Shen, S.; Iosup, A.; and Kuipers, F. 2015. Toxicity detection in multiplayer online games. In *2015 International Workshop on Network and Systems Support for Games (NetGames)*, 1–6.
- Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.
- Nakamura, L. 2015. The unwanted labour of social media: Women of colour call out culture as venture community management. *New Formations* 86(86):106–112.
- Nocentini, A.; Calmaestra, J.; Schultze-Krumbholz, A.; Scheithauer, H.; Ortega, R.; and Menesini, E. 2010. Cyberbullying: Labels, behaviours and definition in three european countries. *Journal of Psychologists and Counsellors in Schools* 20(2):129–142.
- NORC. 2017. New survey: Snapchat and instagram are most popular social media platforms among american teens: Black teens are the most active on social media and messaging apps. *ScienceDaily*.
- Palau, R. M., and Moens, M.-F. 2009. Argumentation mining: the detection, classification and structure of arguments

in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, 98–107. ACM.

Pater, J. A.; Kim, M. K.; Mynatt, E. D.; and Fiesler, C. 2016. Characterizations of online harassment: Comparing policies across social media platforms. In *GROUP*, 369–374. jess-pater.com.

Pfeil, U., and Zaphiris, P. 2007. Patterns of empathy in online communication. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07*, 919. New York, New York, USA: ACM Press.

Phillips, W., and Milner, R. M. 2017. We are all internet bullies. *Time*.

Phillips, W. 2015. *This Is Why We Can't Have Nice Things: Mapping the Relationship between Online Trolling and Mainstream Culture*. The MIT Press.

Preece, J. 1998. Empathic communities: Reaching out across the web. *Interactions* 5(2):32–43.

Reynolds, K.; Kontostathis, A.; and Edwards, L. 2011. Using machine learning to detect cyberbullying. In *10th International Conference on Machine Learning and Applications and Workshops (ICMLA)*, volume 2, 241–244. IEEE.

Roberts, S. 2016. Commercial content moderation: Digital laborers' dirty work. In Noble, S. U., and Tynes, B. M., eds., *The Intersectional Internet: Race, Sex, Class and Culture Online*. New York: Peter Lang Publishing.

Rubin, J. Z.; Pruitt, D. G.; and Kim, S. H. 1994. *Social conflict: Escalation, stalemate, and settlement*. New York: McGraw Hill Book Company, 2nd edition.

Sood, S.; Antin, J.; and Churchill, E. 2012. Profanity use in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '12*, 1481–1490. New York, NY, USA: ACM.

Sood, S. O.; Churchill, E. F.; and Antin, J. 2012. Automatic identification of personal insults on social news sites. *J. Am. Soc. Inf. Sci. Technol.* 63(2):270–285.

Su, X., and Khoshgoftaar, T. M. 2009. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence* 2009:4:2–4:2.

Thompson, N. 2017. The great tech panic: Instagram's kevin systrom wants to clean up the internet. *Wired*.

Wang, L., and Cardie, C. 2016. A piece of my mind: A sentiment analysis approach for online dispute detection. *arXiv preprint arXiv:1606.05704*.

Wang, W.; Chen, L.; Thirunarayan, K.; and Sheth, A. P. 2014. Cursing in english on twitter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 415–425. ACM.

Yang, J., and Leskovec, J. 2011. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 177–186. ACM.

Yin, D.; Xue, Z.; Hong, L.; Davison, B. D.; Kontostathis, A.; and Edwards, L. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB* 2:1–7.