TrollSpot: Detecting misbehavior in commenting platforms

Tai Ching Li
University of California - Riverside
Riverside, CA
tli010@cs.ucr.edu

Joobin Gharibshah
University of California - Riverside
Riverside, CA
jghar002@cs.ucr.edu

Evangelos E. Papalexakis
University of California - Riverside
Riverside, CA
epapalex@cs.ucr.edu

Michalis Faloutsos
University of California - Riverside
Riverside, CA
faloutsos@cs.ucr.edu

ABSTRACT

Commenting platforms, such as Disqus, have emerged as a major online communication platform with millions of users and posts. Their popularity has also attracted parasitic and malicious behaviors, such as trolling and spamming. There has been relatively little research on modeling and safeguarding these platforms. As our key contribution, we develop a systematic approach to detect malicious users on commenting platforms focusing on having: (a) interpretable, and (b) fine-grained classification of malice. Our work has two key novelties: (a) we propose two classifications methods, with one following a two stage approach, which first maps observable features to behaviors and then maps these behaviors to user roles, and (b) we use a comprehensive set of 73 features that span four dimensions of information. We use 7 million comments during a 9 month period, and we show that our classification methods can distinguish between benign, and malicious roles (spammers, trollers, and fanatics) with a 0.904 AUC. Our work is a solid step towards ensuring that commenting platforms are a safe and pleasant medium for the exchange of ideas.

ACM Reference Format:
Tai Ching Li, Joobin Gharibshah, Evangelos E. Papalexakis, and Michalis Faloutsos. 2018. TrollSpot: Detecting misbehavior in commenting platforms. In *Proceedings of WSDM workshop on Misinformation and Misbehavior Mining on the Web (MIS2).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.475/123_4

ACKNOWLEDGMENTS

This work was supported by DHS ST Cyber Security (DDoSD) HSHQDC-14-R-B00017 grant and NSF NeTS 1518878.

1 INTRODUCTION

Any successful medium eventually attracts abusive behaviors, and commenting platforms is no exception. Over the last decade, commenting on news articles has emerged as a new form of highly social interaction. First, a small number of companies facilitate the backend management of comments for a wide range of websites. We use the term **commenting platform** to refer to such platforms, which include Disqus [6], LiveFyre [11], and IntenseDebate [10]. Second, commenting is an intense activity for many users, who can spend many hours daily at it.

We list a set of definitions that we use in this paper. A **user** is defined by a platform account, which enables her to leave comments to articles on a website that uses the commenting platform. A user may leave multiple comments for an article, which leads us to define the **engagement** of a user for that article. An engagement has a time duration and intensity in terms of number of comments. In lack of a better term, when two users comment on the same article, we say that they **collaborate** and we use the term **collaboration** to describe this activity. We use the term **collaboration intensity** referring the number of articles for which two users collaborate.

The key question in our work is: *Can we automatically detect malicious users in these commenting platform?* Specifically, we frame the problem as follows. The input is the commenting information of the users. This includes: the author of the comment, the time it was posted, and information on the article it was posted for. The goal is to identify malicious behaviors and users. Detecting parasitic and abusive behaviors is a critical building block for ensuring that these platforms serve their primary purpose, which is the honest and safe exchange of opinions among readers.

Commenting platforms have attracted little attention so far with only few exceptions [2][1]. Most work on modeling and misbehavior detection focuses on Online Social Networks (OSNs), and blogs. In Section 6, we survey the key related areas including: (a) detecting abusive behaviors and malware propagation in OSNs[3][13]; (b) modeling online user behavior[7] [5] [8]; and (c) analyzing text of online users[15][12][16].

We propose a systematic comprehensive methodology to identify malicious users on commenting platforms to enable: (a) interpretable, and (b) fine-grained classification of malicious behavior. We claim four key novelties in our work.

a. A behavior-based classification. We propose two classification methods, the first one use typical Random Forest classifier yet another one introduces a two-stage classification approach. In this method, we map: (a) observable features to behaviors, and (b) behaviors into user roles, using unsupervised and supervised learning respectively.

b. A comprehensive multidimensional feature set. We combine 73 features from four different dimensions of user interactions: (a) social interaction or user-user interaction, (b) engagement or

MIS2, 2018, Marina Del Rey, CA, USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

user-article interaction, (c) temporal features, and (d) linguistic features.

c. Fine-grained malicious role identification. Our approach goes beyond a good versus bad determination to a more fined-grained classification of misbehaving roles. Here, we focus on three roles: (a) spammers, (b) trolls, and (c) fanatics, which are defined in the next section. However, it is easy to introduce more roles as long as appropriate ground truth is available.

d. Interpretable classification. The results of our approach are interpretable, since they are behavior-centric. A system administrator can better understand or tweak the definition of, say, a spammer, by looking at the behaviors that constitute its definition (e.g. high number of repeated text across articles, low engagement per article).

Our study is grounded on nearly 7 million comments from nearly 200K users over 9 months from Discus, which is arguably the largest commenting platform. Here we highlight some interesting result from our study.

a. Identifying misbehaving users with 0.904 AUC. Our classifier can efficiently identify misbehaving users and it outperforms the baseline we compare with.

b. The role classification result achieves 80.8% overall accuracy.

c. **We find patterns of misbehaving users from their interpretable latent behavior.** An example latent behavior indicates that making longer comments and using capital letters more frequently is a sign of misbehaving users, and the later the active hour of a users, the more likely they are to misbehave.

d. Large scale study. About 0.9% of total users in our data-set are misbehaving users, and CNBC has the largest percentage of misbehaving users which is 1.48%. Only 15% of the misbehaving users exhibit a cross-website behavior.

2 DATA COLLECTION AND DEFINITIONS

Disqus is one of the most widely-used commenting service provider today with a billion unique visitors a month and installed by more than two and a half million sites, including ABC News, Rolling Stone, IGN, Bloomberg and more.

Data Collection. We collected data from Disqus through its Application Programming Interface (API). The API can be used to collect all the comments for a given article, and for all the articles of a given website. However, the functionality of collecting all the comments for a given user would have been useful for our study, but it became unavailable in 2014 for privacy reasons.

Using the website-centric API, we collect data from four popular websites: (a) CNBC News, (b) ABC News, (c) Bloomberg Views and (d) Breaking News - a Disqus channel. The first three are well-known news websites and the last one is the most popular channel on Disqus. A channel is similar to a newsfeed, whose articles are selected by the users that participate in that channel. We collect all comments posted at articles published on these 4 sources in between Nov 1st 2015 and July 31st 2016. The dataset consists of: (a) 286,275 articles, (b) 6,994,693 comments and (c) 201,112 unique users, (d) 1,705,667 engagements.

Establishing the ground truth of misbehaving users. A key challenge in our work is the need for ground truth of misbehaving users. Fortunately, Disqus enables users to report misbehaving comments and the number of reports a comment received is provided by Disqus API. In our 7M comments, we have 66.4k comments reported as "bad/inappropriate" by the community. We will further discuss this in Section 4 and show how we leverage this to identify misbehaving users.

Roles of misbehaving users and ground truth. We identify and attempt to detect three different roles of misbehaving users: trolls, spammers and fanatics. These are inherently difficult to define, so we resort to human feedback.

We define three malicious roles below. We show here the definitions that we gave our evaluators, as we explain in Section 4.

(1) **Trolls.** Users who make inflammatory or inappropriate comments for the sole purpose of upsetting other users and provoking a response.
(2) **Spammers.** Users who repeatedly make similar comments in the same or multiple articles.
(3) **Fanatics.** Users who exhibits an extreme and uncritical enthusiasm in religion or politics.

3 FEATURES AND USER BEHAVIOR

In this section, we identify and study features that relate to user behavior on the Disqus platform. Although studying each of these features in depth is of independent interest, the goal is to identify meaningful features that can help us detect misbehavior. Due to space limitations, we can only provide highlights of the features and their distributions. In Table 1, we outline the features that we use in Section 4.

We study the behavior of users along four dimensions: **(a)** engagement behavior (user-article interaction), **(b)** social behavior (user-user interaction), **(c)** temporal behavior, and **(d)** linguistic properties.

3.1 Engagement behavior

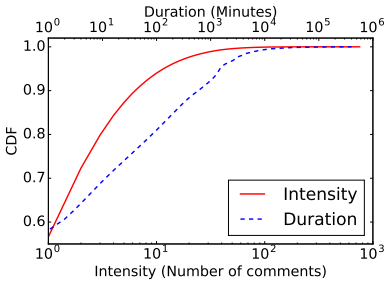


Figure 1: Engagement behavior: Distribution of intensity and duration of engagements.

We quantify the engagement (user - article interaction) with 7 different features. Six of them are derived from two major properties: The **engagement duration** is the time interval between the first comment and the last comment user makes on the article. If the user leaves only one comment, we consider this as zero length interval. The **engagement intensity** is the total number of comments user makes on the article.

Table 1: The overview of the 73 features we use per Dimension

Dimension	Features	Count
Engagement	number of engagements, engagement duration,* engagement intensity*	7
Social	degrees, number of maximal cliques, number of triangles (in different level of collaboration intensity)	17
Temporal	number of comments made in 24-hour slots, highly-active hour	25
Linguistic	number of words,* number of sentences,* percentage of capital letters,* readability metrics, number of URLs	24

* We use several statistical versions (mean, maximum and minimum) of the feature per engagements, user or comment.

Engagement duration: 90% last for less than 10 hours, but some can be as long as half an year. In Figure 1, we plot the Cumulative Distribution Function (CDF) of the duration (top x-axis) for all 1.7M engagements. We find that 90% of engagements last less than 10 hours and have less than 7 comments. Interestingly, we find 106 engagements which last for more than half year!

Engagement intensity: 90% have less than 7 comments, but 0.06% have more than 100 comments. In Figure 1, we plot the CDF of the intensity (bottom x-axis) for all engagements. We find that 90% of them have less than 7 comments, while 1,151 (0.06%) of them have more than 100 comments.

High intensity is correlated with misbehavior. The observations above raise the question: is unusually high engagement intensity associated with misbehavior? Preliminary manual inspection suggests that these engagements contain many reported (by the community) comments. We quantify the correlation between engagement intensity and the total number of reported comments in the engagement and we find a Pearson correlation coefficient $\rho = 0.53$. This suggests that engagement intensity should be a good metric in our classification.

3.2 Social Behavior

We propose 17 features to model the social interaction of the users, which we define as the posting of comments to the same articles.

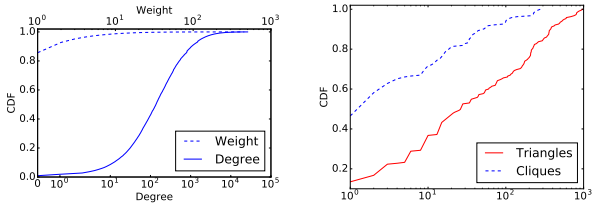
Single-article collaboration Threshold: θ comments. We say that two users collaborate in one article, if they each post at least θ comments on that article. For $\theta = 1$, the graph become very dense, and the analysis is both slow and less informative. In the remaining of this work, we use a threshold $\theta = 2$.

User-user collaboration intensity and threshold: λ articles. The collaboration intensity is the number of articles that two users collaborate for a given threshold θ. To study collaborations at different levels of intensity, we introduce the **collaboration intensity threshold** λ, which we use below.

We define the undirected weighted **collaboration graph** $G_\lambda = \langle V_\lambda, E_\lambda \rangle$ of collaboration intensity λ where:

(1) V_λ is a set nodes v, representing users.
(2) E_λ is the set of edges, where edge e_{ij} between nodes v_i and v_j exists, if and only if the collaboration intensity of the users exceeds the **threshold of λ articles**. The edge weight $w(e_{ij})$ is set to the collaboration intensity.

The collaboration graph (for $\theta = 2$ and $\lambda = 0$) has 95,527 users and roughly 21 millions edges, an average degree of 440.7 and a median degree of 137. Note that we do not include users with zero degree in this graph.



(a) (b)

Figure 2: Social behavior: (a) the CDF of the user degrees (bottom x-axis) and the edge weight distributions (top x-axis), (b) CDF of the number of cliques and triangles of user in the G_{128} collaboration graph.

In Figure 2(a), we plot the CDF of the user degrees (bottom x-axis) and the edge weight distributions (top x-axis). We see that 90% of the users have degree lower than 1,054, the max degree goes to 26,318: this user collaborates with more than 27% of users in the graph! The figure also shows that 90% of the edges have weights less than 2. Although it is not easy to gauge from the plot, we find 12,374 edges with weight over 64, 1,779 edges with weight over 128 and 65 edges with weight over 256. In other words, there are 65 pairs of users who have collaborated on more than 256 articles.

Capturing the collaboration groups: triangles and cliques. We quantify the local connectivity of the users using the number of maximal cliques and triangles in which a user participates. Both these metrics capture how densely connected the neighbors of that user are. Figure 2(b) shows the distribution of triangles and maximal cliques of user in graph, for λ equals to 128. We do not count "trivial" cliques of size two. We see that 50% of users have less than 27 triangles on the neighborhood, 90% of have less than 376, but there are 1% with more than 868 triangles. We also see that 90% of users have less than 51 maximal cliques and there are 8 users, who participate in more than 186 cliques which is more than 50% of all maximal cliques in the graph. These highly collaborative users are suspicious and this encouraged us to consider both these features for misbehavior detection.

3.3 Temporal behavior

We quantify the temporal behavior of users with 25 features.

Most users exhibit persistent behavior: daily and weekly. We use the **time and day of the week** plot to understand the temporal behavior of the users. The plot shows the number of

comments in each hour-of-day and day-of-week as a heat-map. Figure 3(a) shows this behavior for all the users, but individual users exhibit similar daily and weekly behavior, as we will see below. On the x-axis (hour-of-day), we see that the commenting activities increase at the beginning of day in the east coast as annotated in the plot, the peak hours cross the noon period of both east and west coasts then drop sharply after that. Given the news websites we are studying, it is reasonable to assume a US-centric user majority. In y-axis (day-of-week), the activity increases at the start of the week, peaks in the middle, and drops before the weekend. Combining the two observations, we see that Disqus users post most of their comments during the common work-hours in the week. Note that the same pattern is also observed on other social networks, like Facebook [19] and Twitter [17].

The plot lead us to the following hypothesis:

Hypothesis: A typical user makes most of her comments in a period of 3-4 hours in a day.

The highly-active hours of users commenting behavior. To assess the validity of our hypothesis, we would like to answer the question: how many hours a day does a user spend commenting? Determining this hides subtleties as: (a) patterns can vary from day to day, and (b) it could be that many hours have non-zero activity due to averaging over a long period. Thus, we opt to capture how "focused" or "spread" is the commenting activity of the user. We define **highly-active hours** to be the minimum number of hours, during which the user makes more than 50% of their total comments during a day on average.

Highly-active hours is less than 4 hours for 96.7% of the users. The distribution of the users' highly-active hours is shown in Figure 3(b). The plot shows that 96.7% of users have highly-active hours less than four hours, which matches our hypothesis. Interestingly, we find 368 users who have more than 8 highly-active hours, a significantly wider spread. Upon inspection, many of these users have non-trivial activity over 14 hours in a day. This wide range of behaviors suggests that highly-active hours can be a useful metric for our classification.

3.4 Linguistic properties

We identify and study 24 linguistic features from the text of the posts, such as the length of the comment, several metrics readability, number of URLs, but discuss only a two metrics.

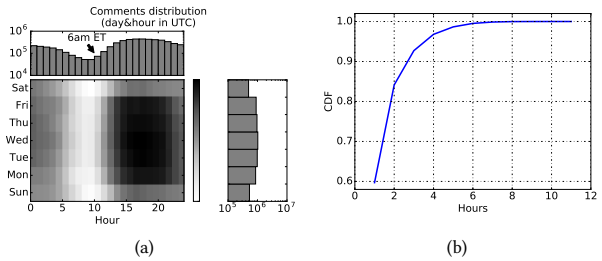


Figure 3: Temporal behavior. (a) Time and day of the week plot. (b) The distribution of highly-active hours of users.

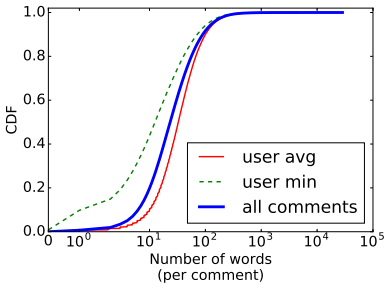


Figure 4: Linguistic property: number of words in a comment

Only 1.9% of comments contain URLs. We naturally consider the existence of URLs in a comment as a feature in our classification: their presence can indicate ad-oriented spamming. We find that only 1.9% of all comments contain one or more URLs. We also find that only 1.7% of the users have ever posted more than 3 comments containing URLs. In fact, our interaction with the data suggests that often users will use URLs as references in support of their opinions.

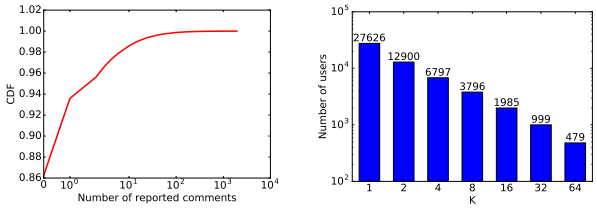
Length: 91% of comments have less than 100 words. The length of a comment is a good metric for the interest and the time a user is willing to spend expressing her opinion about the article. In Figure 4, we plot the distribution of the number of words in a comment in three different ways: **(a)** across all comments, **(b)** the average comment length per user, and **(c)** minimum comment length per user. We see that 91% of comments have less than 100 words, while roughly 20% of the comments with less than 10 words. Interestingly, we also find 14,838 (0.002%) comments with more than 500 words, which is roughly equivalent to 1.5 pages (per wordstopages.com with Times New Roman, 12 font-size, single-spacing): the comment is the size of small article!

Lengthy comments are more likely to be spam. We examine the lengthy comments (>500 words) and find out that 47% of them are verbatim copies of at least one other comment from the same user in our dataset. This is an indication of spamming especially in its broader definition that we adopt here. Such phenomenon is more pronounced for higher word counts: 63% of comments have verbatim copies for length over 1,000 words , and 83% of comments with length 2,000 words or longer. This suggests that comment length is a helpful feature in detecting misbehavior.

4 FEATURE-BASED MISBEHAVIOR IDENTIFICATION

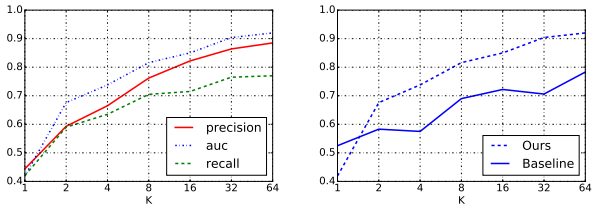
Having identified interesting features, we develop a method to identify misbehaving users and their effectiveness. For convenience, we outline the features in Table 1, and we intend to define them in detail in an extended version of this paper.

A. Establishing the ground truth. To overcome the lack of absolute truth, we rely on "proxy signals". Here, we use the community's own opinion: any user can report (a.k.a. flag) a comment as "inappropriate". Below, we explain how we use this community feedback to construct the ground truth, as the process hides several subtleties.



(a) (b)

Figure 5: (a) Distribution of reported comments of a user. (b) Number of users having k reported comments.



(a) (b)

Figure 6: (a) Classification results as a function of the number of reported comments that "incriminate" a user. (b) AUC of our approach and the baseline.

Ground-truth: Reportings per malicious comment. To increase our confidence, we set a minimum threshold of reports, ϕ, that a comment must have to labeled malicious. The rationale is that a single reporting can be created even accidentally (the authors have regrettably done this once). After analysis and deliberation omitted here, we settled on $\phi = 3$ reportings. We use the term **reported comment** to refer to a comment with more than ϕ reports.

Ground-truth: Reported comments per malicious user. In the same vain, we want to be careful in labeling a user as malicious based on the number of reported comments. We use the **reported comments threshold**, r, to control tune the definition of malicious user.

Roughly 86% of the users have no reported comments. We plot the distribution of the total number of reported comments for each user in Figure 5(a). We find that 86.2% of users have no reported comments, and only 1.6% of users have more than 10 reported comments.

We consider users with zero reported comments as **benign** for the purpose of establishing the ground truth.

Building the ground truth datasets: D_r. We create a set of labeled datasets as follows. First, we distinguish reported users into groups, $R_{r(i)}$, with $r(i) = 2^i, for\ \ i = 0, 1, \ldots$ A user is in group $R_{r(i)}$, if the number of her reported comments are greater or equal $r(i)$. It turns out that no user has more than 128 reported comments. The numbers of users in each group with threshold $r(i)$ are shown in Figure 5(b). Second, we create datasets $D_{r(i)}$ by randomly selecting 200 reported users from $R_{r(i)}$, and combining them with 200 benign users (zero reported comment).

Benign user selection: Minimizing the effect of the number of comments of a user. When selecting the 200 benign users to build the datasets, we follow the insightful approach of the earlier in this space [2]. Namely, we find benign users that "match" the number of comments of reported users. The goal is to minimize the dominant role that the number of comments of a user can play. First, the number of comments and number of reported comments of a user are strongly positively correlated with a Pearson Correlation Coefficient of 0.73. Second, most users have few comments given the skewness of the distribution. Thus, a purely random selection of benign users would have probably created a low-activity benign set in terms of user comments, and a highly-active reported users. The classification would have been very accurate by relying heavily on the number posts.

B. The benign-malicious classification. For the classification, we use the Random Forest classifier provided by Weka [20], which gave the best results among many that we tried. We perform ten-fold cross validation and report the precision, recall, and **ROC curve (AUC)** of each dataset in Figure 6(a). The plot shows that our features can identify reported users with more than 80% precision when the threshold r is larger than 16.

Selecting reported comments threshold $r = 32$. We manually examined reported users in D_{16}, D_{32} and D_{64} by sampling 20 users from each group. We find that users with more than 16 reported comments exhibit a persistent misbehavior throughout their lifetime. The 40 reported users sampled from D_{32} and D_{64} are 100% labeled as misbehaving users by our independent human evaluators. This gives us confidence to claim that users with 32 or more reported comments are misbehaving users. Thus, we use dataset D_{32} as reference below.

The baseline classifier. We adapt and use a previously proposed algorithm as a reference in our study [2]. That work focuses on predicting whether a user will get banned in the future, which is a somewhat different goal from our work.

We faithfully reproduce the reference method [2], which also uses the Random Forest technique like we do. In their work, they use 35 features, however 6 are not publicly available, namely the website moderator features, such as the number of up-votes given to others (similar to a Facebook "Like"). Unfortunately, we were not able to access those features through the public API, therefore we do not use them in our experiments.

The accuracy of our method exhibits 90% AUC. For clarity of presentation, we only show the comparison of AUC of the two methods in Figure 6(b). Our method performs better than the baseline classifier, with the caveat that both approaches are restricted to the publicly available features. The most interesting conclusion is that our features have good discriminatory power in classifying misbehaving and benign users.

C. The fine-grained malicious role classification. The novelty of the work relies partly on going beyond the "malicious" label to the role of the misbehaving user.

Ground truth for the roles of misbehaving users. Here, we resort to manual labeling to create our reference data: we examine 200 misbehaving users in D_{32} and categorize them into three different roles: trolls, spammers and fanatics. To the best of our

Table 2: Role classification result

Role	Precision	Recall
Trolls	86.4%	73.1%
Spammers	58.6%	81%
Fanatics	86.1%	73.3%
Benign	81%	87.5%
Overall accuracy	80.8%	

Table 3: Misbehaving users in websites

Website	Total users	Misbehaving users
ABC News	129,053	717 (0.56%)
CNBC	78,618	1167 (1.48%)
Bloomberg Views	33,223	131 (0.39%)
Breaking-News (Disqus channel)	11,607	48 (0.41%)

knowledge, this is the first work that does such a fine-grained classification of user behaviors, especially in commenting platforms.

Each user is labeled by three evaluators who were presented with the definitions in Section 2. In absence of unanimity, we set their role by taking the majority label. Out of 200 misbehaving users, 104 are labeled as trolls, 21 as spammers and 75 as fanatics.

Role classification: Our approach has 80.8% overall accuracy. Having ground truth, we apply 10-fold cross validation with the same classifier and the same 73 features that we used to identify misbehaving users. Our result shows that our model can effectively classify the role of misbehaving users with an overall accuracy of 80.8% as shown in Table 2. For all the classes the recall is above 73% and the precision above 81% except the Spammers.

The community shows tolerance to non-provocative spammers. Intrigued by the low precision for spammers, we find that spam comments without provocative language, swear words and sarcasm often do not get reported by the community. Recalling Section 3.4, the unusually long comments are more likely to be spam, due to verbatim repetitions. However, this behaviors seems to either escape detection or be met with tolerance.

Our method identifies un-reported spammers. We examine the 12 false positive in the spam category: our spam label is not corroborated by the community. We actually find that at least one of these users exhibits clear spamming behavior, as she repeats the exact same comment 3 times in one article and 5 times in another. This investigation suggests that our approach could be catching parasitic behaviors that the community could miss. As a result, the accuracy of our algorithm could be better than reported here, especially for spammers.

D. Studying the misbehaving users in the wild.

We apply the classier to the whole dataset to understand how misbehaving users distributed in the wild. For identifying misbehaving users, the classifier labeled 1,738 (0.86%) users as misbehaving with confidence over 80%. The role classification results are: 866 users are labeled as trolls, 165 users are labeled as spammers and 597 users are labeled as fanatics. We also study how misbehaving users distributed in each websites, the result is shown in Table 3. It shows that CNBC has the largest number of misbehaving users with 1,167 which is 1.48% of total users active on the website while the rest of websites are at most 0.56% of users are misbehaving users. Again, we only label misbehaving users with confidence over 80%.

5 INTERPRETABLE TWO-STAGE CLASSIFICATION

In this section, we introduce a novel, two-stage approach in classifying misbehaving users. In the first step we identify tightly-knit groups of users who exhibit very similar behavior across a certain subset of the features we discussed in the previous section. Each such group of users and their associated features defines a *latent user behavior*. Note that we do not define those latent user behaviors a-priori, they rather emerge from the data automatically. Subsequently, we use these latent user behaviors as our new features towards classifying misbehaving users. The advantage of this two-stage approach is that the latent user behaviors offer interpretability of the results and allow for in-depth analysis of different user (mis-)behaviors.

5.1 Identifying latent behaviors: Co-clustering

The first stage of our two-stage approach can be exactly mapped to an instance of *co-clustering*. In a nutshell, co-clustering is the simultaneous clustering of all data modalities. In our case, consider a representation of a user in the vector space defined by the entire set of their features. Hence, the modalities of our data are two: "user" and "feature". Co-clustering of a users $\times$ features data matrix essentially yields different groups (or co-clusters), where each one contains a subset of the users and a subset of the features. In other words, when grouping users and features together, co-clustering allows for the flexibility of finding a set of users that are highly similar *only for a subset* of their features. This subset of features is key to our method. Essentially, the subset of features of a co-cluster defines a **latent user behavior**.

An illustrative example is shown in Figure 7, where we have four distinct latent user behaviors: "Benign", "Troll", "Spammer", and "Fanatic". After properly rearranging the users and the features, in this simple example we see that each latent user behavior is forming a block within the matrix. Notice that blocks can overlap since: (a) in some cases a user can be classified with different types of behavior, and (b) some features may be associated with multiple types of behavior. Mathematically, each block (formally co-cluster) in the data can be seen as (approximately) a rank-one matrix. By definition, a rank-one matrix can be written as the outer product of two vectors, i.e., $\mathbf{a}\mathbf{b}^T$. We, thus, can use the following formulation for co-clustering, which was originally proposed in [14]:

$$\min_{\mathbf{A}\geq\mathbf{0},\mathbf{B}\geq\mathbf{0}} \|\mathbf{X} - \mathbf{A}\mathbf{B}^T\| + \lambda \sum_{i,r} |\mathbf{A}(i,r)| + \lambda \sum_{j,r} |\mathbf{B}(j,r)| \qquad (1)$$

The above equation decomposes the data matrix $\mathbf{X}$ into a sum of rank-one matrices (each one being approximately a co-cluster) and

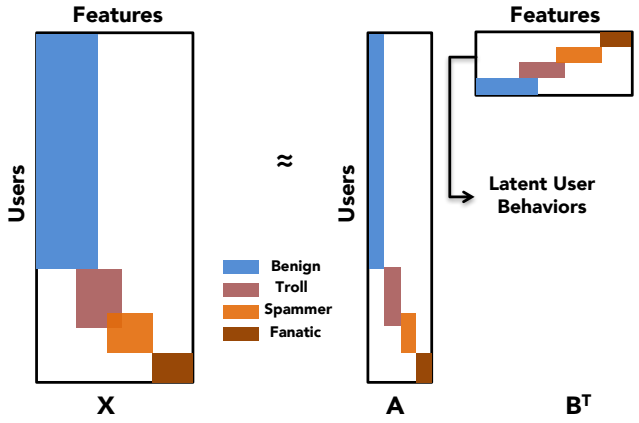


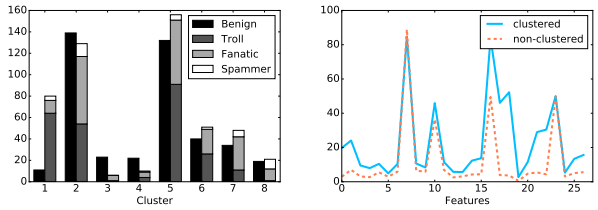
Figure 7: Co-clustering discovers latent user behaviors and the assignment of users and features to those behaviors.

the regularization penalties on the ℓ_1 norms of matrices **A**, **B** (whose columns hold the indicator vector for each co-cluster, as shown in Fig. 7) ensure that the solution is sparse, so that only the users and features that belong to a particular co-cluster have a non-zero (and positive) value in our indicator vectors. Using this type of overlapping co-clustering, a user may be assigned to multiple latent behaviors at the same time, with varying degrees of participation. This enhances interpretability because it can give further insight on whether e.g., a user has more than one roles or their account has been hijacked and some of their comments are benign but some are malicious. Finally, the above co-clustering formulation is *lossy*, which means that it does not require all users to belong to at least one co-cluster. In other words, the above formulation focuses on identifying the most dominant latent behaviors and the users that engage in them, and leaves out users who do not clearly belong in one (or more) of those latent behaviors. We, henceforth, refer to those users as "non-clustered".

5.2 Methodology

We apply co-clustering to the same dataset we used for role classification. This dataset contains 200 benign users and 200 misbehaving users, which include 104 trolls, 21 spammers and 75 fanatics. After fine tuning the parameters of co-clustering, we find that the best number of co-clusters in the data is 8. An indication for this number is that the algorithm returned empty clusters for larger numbers of clusters. We show the distribution of users in each clusters in Figure 8(a).

The size of clusters varies from 29 to 268, indicating that co-clustering could not only capture the common behaviors but also the obscure ones (small clusters). In each cluster, the clustered users will exhibit a different level of value on multiple features comparing to non-clustered users as we present in Figure 8(b). The X-axis demonstrates the 27 features used to capture the cluster and the Y-axis is the mean of features value. Clustered users have a much higher value on feature 18 and 21 which are the percentage of capital letters and number of sentences. This explains users in this cluster exhibits a *latent behavior* of using more capital letters and more sentences in their comments.



(a) (b)

Figure 8: Co-clustering. (a) Distribution of users in each cluster. (b) Example of latent behavior from cluster 1.

Using latent behaviors to identify misbehaving users: 0.77 AUC. In order to show that the latent behaviors can be used to identify misbehaving users, we take **A** matrix from the co-clustering result. Each row in the matrix is a user and each column corresponds to a cluster, with the (i, j) value being the "membership" of user i to the cluster j. We use these values as features to train a classifier. Note that for each set of features we use the classifier that yielded the highest performance. In the case of raw features this was Random Forest, but for the case of the latent behavior features it was Support Vector Machine (SVM) [1]. We thus report results using SVM. Our result shows that we can still accurately identify misbehaving users with an AUC of 0.77 with 86.3% precision and 63.5% recall. The role classification result is shown in Table 4.

Using latent behaviors to identify misbehaving roles: 73.3% overall accuracy. Using 10-fold cross validation, we assess the ability of our method to identify misbehaving roles, and we show the results in Table 4.

The overall classification accuracy when using the latent behaviors is lower than using the raw features, albeit still very good. However, we argue that the reduction in accuracy comes with an increase in the *interpretability* of the classification using latent behaviors, as we present below.

Interpreting latent behaviors. Due to the space limitation, we are not able to discuss all latent behaviors in the data. We will, thus, focus on the latent behaviors that help us understand misbehaving users.

(1) Latent behavior 1: Users who exhibit such behavior are, on average, making more comments to articles, are highly active between 7-13 UTC (midnight in US), use more capital letters and also more characters and sentences on average in their comments.
(2) Latent behavior 3: Users are highly active between 13-26 UTC, having ~60% more comments and degrees of collaboration than the average.
(3) Latent behavior 4: Users are active between 3-8 UTC, making slightly more (24%) comments than the average.

[1]This behavior is expected since the latent behavior features, contrary to the raw features, are embeddings of the users in a vector space, thus SVM should work better than Random Forest.

Table 4: Role classification result (co-clustering)

Role	Precision	Recall
Trolls	78.7%	71.2%
Spammers	60.7%	81%
Fanatics	76.4%	73.3%
Benign	71.4%	73.5%
Overall accuracy	73.3%	

(4) Latent behavior 6: Users are active between 6-14 UTC, making comments which have better readability.

Figure 8(a) shows that 87.9% of users who exhibit latent behavior 1 are misbehaving users. This is in contrast to users of latent behavior 6, which are active at similar periods of time but only 56% of them are misbehaving users. This explains that using more capital letters and making longer comments are the characteristic of misbehaving users.

Another interesting observation here is that, the active period could also imply whether a user is misbehaving or not. We see that there is an active time shift from latent behavior 3, latent behavior 4 to latent behavior 6, and the percentage of misbehaving users are 20.7%, 31.3% and 56% respectively. If we assume that most of the users are active in the timezone between ET and PT, then we may further deduce that the later at night a user is active, the more likely they are to be misbehaving.

6 RELATED WORK

We only have space to review indicative studies in five areas.

Detecting malicious behavior. Many studies detect anti-social behavior [2][1], bots[3][13] or cyber-bullying [9] but they focus on OSNs. A recent work [2] analyzes the comments on three websites using Disqus, but the focus is to predict if a user will be banned due to misbehavior, which is somewhat different from ours: (a) some misbehaviors are not necessarily reported, as we saw, and (b) we have a fine-grained definition of misbehavior.

Detecting malicious behavior. Many studies detect anti-social behavior [2][1], bots[3][13] or cyber-bullying [9] but they focus on OSNs. A recent work [2] analyzes the comments on three websites using Disqus, but the focus is to predict if a user will be banned due to misbehavior, which is somewhat different from ours: (a) some misbehaviors are not necessarily reported, as we saw, and (b) we have a fine-grained definition of misbehavior.

Analyzing text of online users. Many studies leverage textual analysis to detect spammers on OSNs [15] and blogs [12] or how to achieve the same objective by analyzing the behavioral patterns [16].

Modeling and understanding online user behavior. Several works study temporal and behavioral patterns of OSN users [7] [5] [8]. Some recent studies also use mobile phones and OSNs to profile users' psychological states [18][4].

7 CONCLUSION

We develop a systematic and comprehensive methodology to identify malicious users on commenting platforms. The novelty of our work lies in its: (a) fine-grained classification of malicious behavior, and (b) interpretable approach to classification. From an algorithmic point of view our work uses a comprehensive set of 73 features across four different types of information, and a novel two-stage classification.

The overall classification accuracy of our approach is 80.8% for the fine-grained 4-class problem. We also identify several unusual and surprising behaviors, and we provide a broader understanding of the users of these platforms.

Our work is a first step towards an efficient, easy-to-understand and easy-to-manage approach to safeguarding commenting platforms from parasitic and malicious users.

In the future, we aim at: (a) create a public larger labeled datasets to facilitate further research, (b) consider more categories of malicious roles, and (c) conduct large scale study of the evolution of malicious behavior on commenting platforms.

REFERENCES

[1] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *CSCW*.
[2] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial Behavior in Online Discussion Communities. *arXiv preprint arXiv:1504.00680* (2015).
[3] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. 2012. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? 9, 6 (2012), 811–824.
[4] Munmun De Choudhury, Scott Counts, Eric J Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *CSCW*. ACM, 626–638.
[5] Pravallika Devineni, Danai Koutra, Michalis Faloutsos, and Christos Faloutsos. 2015. If walls could talk: Patterns and anomalies in Facebook wallposts. In *ASONAM*. ACM, 367–374.
[6] Disqus. 2015. Disqus: Blog-comment hosting service. https://disqus.com. (2015).
[7] Alceu Ferraz Costa, Yuto Yamaguchi, Agma Juci Machado Traina, Caetano Traina Jr, and Christos Faloutsos. 2015. Rsc: Mining and modeling temporal activity in social media. In *SIGKDD*. ACM, 269–278.
[8] Xiaotao Gu, Hong Yang, Jie Tang, and Jing Zhang. 2016. Web user profiling using data redundancy. In *ASONAM*. IEEE, 358–365.
[9] Homa Hosseinmardi, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2016. Prediction of cyberbullying incidents in a media-based social network. In *ASONAM*. IEEE, 186–192.
[10] Intense Debate. [n. d.]. Intense Debate: Imagine better comments. http://intensedebate.com. ([n. d.]).
[11] LiveFyre. [n. d.]. LiveFyre: Real-time Content Marketing and Engagement. http://web.livefyre.com. ([n. d.]).
[12] Gilad Mishne, David Carmel, and Ronny Lempel. 2005. Blocking Blog Spam with Language Model Disagreement.. In *AIRWeb*, Vol. 5. 1–6.
[13] Fred Morstatter, Liang Wu, Tahora H Nazer, Kathleen M Carley, and Huan Liu. 2016. A new approach to bot detection: striking the balance between precision and recall. In *ASONAM*. IEEE, 533–540.
[14] Evangelos E Papalexakis, Nicholas D Sidiropoulos, and Rasmus Bro. 2013. From k-means to higher-way co-clustering: Multilinear decomposition with sparse latent factors. 61, 2 (2013), 493–506.
[15] D Sculley and Gabriel M Wachman. 2007. Relaxed online SVMs for spam filtering. In *SIGIR*. ACM, 415–422.
[16] Ashish Sureka. 2011. Mining user comment activity for detecting forum spammers in youtube. *arXiv preprint arXiv:1103.5044* (2011).
[17] Sysomos. 2014. Inside Twitter: An In-Depth Look Inside the Twitter World. (April 2014). http://sysomos.com/sites/default/files/Inside-Twitter-BySysomos.pdf
[18] Rui Wang et al. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *UbiComp*. ACM, 3–14.
[19] Christina Warren. 2010. When Are Facebook Users Most Active? [STUDY]. (2010). http://mashable.com/2010/10/28/facebook-activity-study/
[20] Ian H Witten, Eibe Frank, Leonard E Trigg, Mark A Hall, Geoffrey Holmes, and Sally Jo Cunningham. 1999. Weka: Practical machine learning tools and techniques with Java implementations. (1999).