The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data

Eshwar Chandrasekharan[1], **Mattia Samory**[2], **Anirudh Srinivasan**[1], **Eric Gilbert**[1]
[1]Georgia Institute of Technology, [2]University of Padua
[1]Atlanta GA 30332 USA, [2]35122 Padua Italy
eshwar3@gatech.edu, samoryma@dei.unipd.it, asrinivasan45@gatech.edu, gilbert@cc.gatech.edu

ABSTRACT
Since its earliest days, harassment and abuse have plagued the Internet. Recent research has focused on in-domain methods to detect abusive content and faces several challenges, most notably the need to obtain large training corpora. In this paper, we introduce a novel computational approach to address this problem called *Bag of Communities* (BoC)—a technique that leverages large-scale, preexisting data from other Internet communities. We then apply BoC toward identifying abusive behavior within a major Internet community. Specifically, we compute a post's similarity to 9 other communities from 4chan, Reddit, Voat and MetaFilter. We show that a BoC model can be used on communities "off the shelf" with roughly 75% accuracy—no training examples are needed from the target community. A dynamic BoC model achieves 91.18% accuracy after seeing 100,000 human-moderated posts, and uniformly outperforms in-domain methods. Using this conceptual and empirical work, we argue that the BoC approach may allow communities to deal with a range of common problems, like abusive behavior, faster and with fewer engineering resources.

Author Keywords
social computing; online communities; abusive behavior; moderation; machine learning

ACM Classification Keywords
H.4.m. Information Systems Applications: Miscellaneous

INTRODUCTION
A key challenge for online communities is moderation. For example, the founders of the social media startup Yik Yak spent months of their early time removing hate speech [6]. Twitter has stated publicly that dealing with abusive behavior remains its most pressing challenge [63]. Many sites have disabled the ability to comment at all because of problems moderating those spaces [15], and empirical work has shown that people leave platforms after being the victims of online abuse [27]. Moreover, recent Pew surveys indicate that abuse happens online much more frequently than often suspected: approximately 40% of Internet users report being the subject of online abuse at some point, with underrepresented users most often targeted [18, 19, 25].

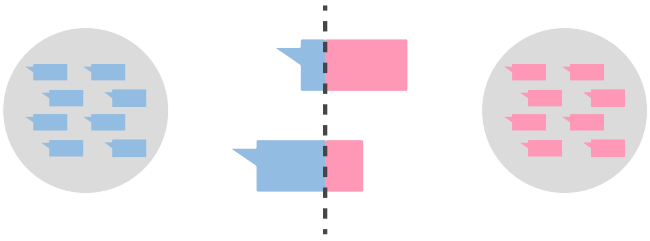

Figure 1. A conceptual illustration of Bag of Communities approach, here with two source communities employed. When new and unlabeled posts are generated in a community, similarity scores can be assigned by comparing them to preexisting posts from other communities (*blue* and *pink*, in this example). A downstream classifier uses similarity scores to make predictions, in our case about abusive behavior.

On most sites today, moderation takes two primary forms: distributed social moderation [20, 30, 49, 50] and machine learning-based approaches [6, 10]. In the former, a site's users triage submissions via voting or reporting mechanisms—after which the site can take action. In the latter, online communities compile large datasets of example posts that have been moderated off-site, and thereafter train machine learning algorithms. The distributed social moderation approach is appealing because it can be deployed quickly and easily, and offloads the work of moderation to a large human workforce; yet, it requires vast amounts of human labor from the very people you would rather not see abusive posts (i.e., your users). Machine learning-based approaches can help by algorithmically triaging comments for a much smaller number of (perhaps paid) human moderators; yet, they typically require vast amounts of labeled training data.

This paper bridges this data gap by introducing a new analytic concept for studying and building online communities: the *Bag of Communities* (BoC) approach. In brief, BoC aims to sidestep site-specific models (and their data) by computing similarity scores between one community's data and preexisting data from other online communities. In this paper, we in-

CHI 2017, May 06 - 11, 2017, Denver, CO, USA

ACM 978-1-4503-4655-9/17/05$15.00
DOI: http://dx.doi.org/10.1145/3025453.3026018

troduce the concept of BoC and use it in an "existence proof:" identifying abusive posts from a major online community. Relying on 10M posts collected from 9 different online communities from 4chan, Reddit, Voat and MetaFilter, we show that a BoC-based linguistic classifier outperforms an in-domain classifier with access to over 4 years of site-specific data. We demonstrate that a BoC classifier can be used on a target community[1] off the shelf with roughly 75% accuracy—no training examples are needed from the target community. That is, an algorithm with access to only *out-of-domain* data can predict abusive posts in another community—without access to data from that community. In addition to this static model, we also explore a dynamic BoC model mimicking scenarios where newly moderated data arrives in batches. It outperforms a solely in-domain model at every batch size, achieving 91.18% accuracy (95% precision) after seeing 100,000 human-moderated posts. This is notable since it implies that while the BoC approach will help communities without moderators to generate training data (static model), the BoC will continue to boost systems that predict abusive behavior after years of professional moderation (dynamic model).

For CMC and HCI theory, BoC provides a new analytic lens through which existing online phenomena may be examined. For example, a researcher might use BoC to empirically derive a taxonomy of online communities based on their similarity to one another. From a systems perspective, BoC may allow sites to address a variety of common problems. In addition to identifying abusive behavior (the focus of the present work), sites need to sort content based on its likelihood to be interacted with, identify spam, and decide whether a post requires intervention by professionals (e.g., suicidal ideation). In the latter case, for example, one could imagine deploying BoC using r/suicidewatch, a Reddit suicide support forum, as a companion data source. In essence, BoC could allow communities (especially new ones with limited resources) to spend their time on what *differentiates* them from other places on the Internet, and less time on common problems shared across sites.

Next, we review existing work, and then formally define BoC. Then, we present our empirical investigation into identifying abusive posts from a major online community using data gathered from 4chan, Reddit, Voat and MetaFilter. Finally, we reflect on some of the opportunities presented by BoC, as well as some of the limitations and opportunities introduced by our empirical work.

RELATED WORK

In this section, we discuss related work on online antisocial behavior, and how it often focuses on in-domain methods. We conclude by laying out the challenges faced by current methods, and discuss how our work helps address these problems.

Commonly deployed moderation approaches

Having plagued online communities for years, technical, design and moderation approaches have been invented to cope with abusive posts (e.g., [28, 31, 34]). A simple existing approach is to designate a separate place where members can "take it outside," places sometimes communicated through FAQs (such as on Usenet) [56]. Reprimands for violations might include private emails, or even public censure [56]. To take one example, many sports-centric forums have designated trash-talk threads, suggesting to members that the behavior is acceptable here but not elsewhere. More sophisticated approaches exist as well. In widespread use today is distributed social moderation, on sites such as Reddit, Hacker News, Yahoo! Answers, Facebook, Yik Yak and Slashdot [20, 30, 49, 50]. In this model, other users vote up or vote down content as they see fit, perhaps even reporting highly objectionable content through special reporting mechanisms (i.e., Facebook's bullying report mechanism, or Reddit's reporting mechanisms). Also in widespread use are centralized moderation mechanisms, embodied in technical designs on sites like Reddit and other forums, where a small number of power users maintain order over the community by removing abusive posts manually. Several popular sites, like YouTube and Facebook, have teams of human moderators, who manually go through posts, and scrub the site of offensive or malicious content [7]. Finally, a handful of slightly more technical approaches are reported to exist within certain communities: word-ban lists and source-ban lists take action when users try to post something objectionable, or from somewhere objectionable. For example, sites like Yik Yak employ manually assembled word-ban lists (i.e., specific terms that flag a user's post) [6], and sites like Hacker News and Yelp employ source-ban lists where users may not post from IP addresses originating from known proxies and Tor. Some communities take the approach of promoting quality content, rather than demoting abuse, which might have the desired effect of improving online discourse. Studies have investigated the extent to which a subset of the criteria at play in the selection of high quality comments by the New York Times as "NYT Picks," can be operationalized computationally [14, 42].

Drawbacks of deployed approaches

While partially effective and used at scale, all the existing approaches described above suffer from drawbacks. First, the "take it outside" approach assumes that the objectionable behavior will remain locked away in the specially-designated area; however, we know from press accounts [41] as well as academic research (e.g., [40]) that norms often bleed over into neighboring communities. Second, both the distributed and centralized moderation approaches require a great deal of human labor [29, 44, 64]. In the centralized approach, the labor falls on a small number of volunteers who must work tirelessly to maintain the community; in the distributed approach, sites ask their users to deal with exactly the type of content they wish their users did not have to see. Moreover, we know from empirical work that these voting mechanisms are susceptible to herding effects [36], underprovision [21], and potential collusion as flagging can be used to indicate disagreement or dislike of a post that is not otherwise inappropriate or profane [32], casting doubt on their reliability. Third, the more technical word-ban lists and source-ban lists are crude by modern standards, and have been observed to perform poorly [57]. For word-ban lists, you need only consider in how many different ways swear words are used (i.e.,

[1] Anonymized for reasons explained later.

exclamation, disbelief, exasperation, insult, etc.) to see the difficulty in applying blanket word-bans. Systems that only consider the source (i.e., banning or shadow-banning users coming from Tor [17]) inadvertently censor users who, for example, try to hide from repressive governments.

In-domain approaches to moderate antisocial behavior

Prior research has looked at technical approaches to moderating online antisocial behavior. All of them tend to focus on in-domain methods to study different kinds of antisocial behavior and develop strategies to counter them. Studies have shown that antisocial behavior like undesirable posting [9, 10, 59], and textual cyberbullying [16, 65] can be identified based on the presence of insults, user behavior and topic models. In recent years, politeness has been studied in online settings, to help keep online interactions more civil. Researchers have built a *politeness* classifier using a computational framework for identifying linguistic aspects of politeness [11]. While these methods are effective within-domain, learning *across domains or communities* remains an open question.

Challenges faced by current work

Current moderation techniques employed by researchers and community moderators face key challenges. Supervised detection techniques require labeled ground truth data for building and evaluating a model. These data are difficult to obtain, and manual annotation is a common approach to address this challenge. But this task requires a large amount of manual labor to hand-annotate (or label) the data. This method is also inherently subject to biases in the annotator's judgment, which could affect the quality of the analysis results [8].

A constant struggle is to identify good data sets that researchers can study. Communities do not publicly share data containing moderated content due to privacy and public relations concerns. This restricts data access, and makes it difficult to model the types of abuse present in a community. In addition, new and emerging online communities lack enough data from their respective users. A new community has by definition few contributions, and therefore even fewer labeled examples; this does not allow them to build robust automated detection systems for identifying abusive content. As a result, building cross-domain moderation systems remains a challenge. Yet, studies have shown that it is important to define community tolerance for abusive behavior as early as possible [61].

BoC aims to address many, but not all, of these challenges. In particular, cross-community similarity allows online communities to piggyback on the data of others, requiring far fewer (and perhaps no) labeled training examples. BoC may form the backbone of cross-domain classifiers built on the data of many Internet communities.

BAG OF COMMUNITIES (BOC)

In this section, we define a new approach to identify certain kinds of online behavior by leveraging large-scale, preexisting data from other Internet communities. The intuition behind our approach is to use the similarity of a post to a known, existing community as a feature in later classification. For example, a post that seems at home within a corpus of 4chan posts may likely be inappropriate for npr.org.

First, we define a method to compute cross-community similarity (CCS), a building block of our approach. We then introduce a new model where a variety of CCS data points act in concert to aid predictions in a new community. Analogous to the well-known Bag of Words representation, we call this the *Bag of Communities* approach.

Cross-Community Similarity (CCS)

Let S be a source community, with whose data we will compare a community of interest—or target community T. While one could approach representing S and T in a variety of ways, it seems natural to model S and T via their posts: let $p \in \mathbb{R}^n$ be a vector-space representation of a post in n dimensions. S and T then comprise all vectors corresponding to their constituent posts. One dimension might represent whether the post was created on a weekday, another might represent whether it contains the word "happy," another might represent whether the post contains an image, etc.

S and T could represent posts along a variety of (possibly infinite) dimensions: temporal characteristics (burstiness vs. spread-out), posting medium (textual vs. image-centric), network structure (connected vs. disconnected), identity (anonymous vs. identifiable), community norms (supportive vs. judgmental). In this paper, we focus on a linguistic representation: the words and phrases used in S and T serve to define S and T. That is to say, a post is represented as a vector with 1's connoting a word or phrase's presence, and 0's otherwise.

There are as many ways to compute $CCS(S,T)$ as there are to compute similarity between vector spaces [1, 24, 33]. Its application may drive the particular method. For example, a straightforward approach might involve computing the centroids s and t of S and T, respectively, and next computing $cos(\theta)$ for the angle θ between them. However, we adopt an approach in this paper inspired by Granger causality [23]. Let M_S be a statistical model that predicts (real-valued) membership in S. $CCS(S,T)$ is then the information provided by $M_S(p)$ in predicting membership in T, for some post p. In other words, we let a model predicting membership in S to predict membership in T. This is analogous to the Granger-causal idea of letting one time series at time t predict the value in another time series at time $t+k$. By "information provided by $M_S(p)$," we mean that M_S may not be used directly, but as the raw material of some encapsulating function. The range of $CCS(S,T)$ is $[0,1]$, with $CCS(S,T) = 0$ for entirely dissimilar communities and $CCS(S,T) = 1$ for entirely similar communities.

Bag of Communities definition

In a Bag of Communities representation, a post $p \in T$ generates CCS scores $CCS(S_1,T)$, $CCS(S_2,T)$, $CCS(S_3,T)$, ... for a variety of source communities S_1, S_2, S_3, ... A Bag of Communities model develops a function $f(CCS(S_1,T), CCS(S_2,T), CCS(S_3,T), ..., T)$ that maps these CCS scores and local, site-specific information to a prediction in $[0,1]$.

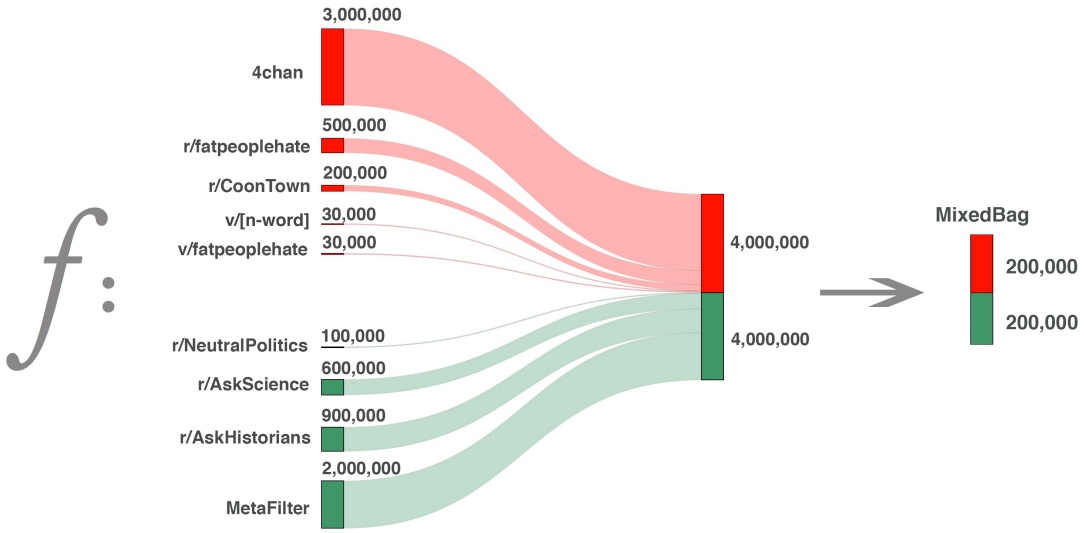

Figure 2. An illustration of the overarching Bag of Communities concept, along with the approximate number of posts collected from each source community in our empirical work. Cross-community similarity values are obtained by comparing target community posts to preexisting posts from source communities ($CCS(S_i, T)$). Communities in *red* were selected because we hypothesized they contained abusive content and those in *green* because they are well-moderated. The goal is to learn a function that maps the source communities to the target community.

In other words, as illustrated in Figure 1, a post might be compared against 4chan, MetaFilter, hateful subreddits, etc. These scores are then fed to another, higher-level classifier that also takes site-specific information into account. In other words, an ensemble classifier might use site-specific information (e.g., the words and phrases used in *that* community) along with CCS scores representing similarity to 4chan, MetaFilter, etc. to make a prediction. Figure 2 illustrates the overall process in function notation, mapping the domain of source communities to the range of the target community.

SOURCE AND TARGET COMMUNITIES

Next, we explore using the BoC approach in predicting abusive behavior in a new online community. We draw data from 9 communities from 4chan, Reddit, MetaFilter and Voat—with MixedBag[2] as our *target* community. Our BoC models will aim to learn from content on source communities, and make predictions about a post's likelihood of being labeled as abusive on MixedBag.

Next, we look at our source and target communities in more detail, and explain the motivation behind using each community to build our BoC models.

Source: 4chan's /b/ and /pol/

4chan is made up of themed online discussion boards, where users generally post anonymously. 4chan is popularly known as the "Internet hate machine" [53], and "the rude, raunchy, underbelly of the Internet" [39]. The use of racist, sexist and homophobic language is common on 4chan. Groups are often referred to using a "fag" suffix (e.g., new members are "newfags", British users are "britfags"), and a common response to any self-shot picture by a woman is "tits or GTFO" [3].

/pol/ is 4chan's *politically incorrect* board. As per 4chan's rules page, /pol/ is a board where debate and discussion related to politics and current events is welcome. /b/ is 4chan's "random" board, and is 4chan's first and most active board, representing 30% of all 4chan traffic. In the words of its creator, /b/ is the "life force of the website," and a place for "rowdiness and lawlessness" [60]. These boards are infamous for exhibiting a range of explicit content. Despite being a funny, open and creative board that is credited for the creation and promotion of numerous memes, the content on /b/ is frequently intentionally offensive, with little held sacred.

Source: Reddit's r/fatpeoplehate and r/CoonTown

In the wake of Reddit's new anti-harassment policy, the website banned several hate communities that it found in violation of the site's rules [4, 52]. According to Reddit's announcement, "We will ban subreddits that allow their communities to use the subreddit as a platform to harass individuals when moderators don't take action." [48]

We collected posts from two of Reddit's most controversial communities which routinely engaged in hate speech,

[2]Pseudonym for the actual target website/community, as per research agreement.

namely *r/fatpeoplehate* and *r/CoonTown*. r/fatpeoplehate is a fat shaming community devoted to posting (among other things) pictures of overweight people for ridicule [52]. It was one of the most prominent removals from Reddit, and had 151,404 subscribers at the time of its banning, as reported by Reddit Metrics.[3]

r/CoonTown is a racist subreddit dedicated to violent hate speech against black people. It contained "a buffet of crude jokes and racial slurs, complaints about the liberal media, links to news stories that highlight black-on-white crime or Confederate pride, and discussions of black people appropriating white culture." [35] It had 21,168 subscribers at the time of banning, as reported by Reddit Metrics.[4]

Source: Voat's v/fatpeoplehate and v/[n-word]

Voat is a media aggregator website which claims to emphasize free speech above all other values. Following Reddit's banning of subreddits for violating its harassment policy, users from those banned communities migrated to Voat, creating hate subverses to take the place of their banned subreddit counterparts [26, 37]. In particular, we collected posts from v/[n-word] and v/fatpeoplehate, which are the Voat equivalents of r/CoonTown and r/fatpeoplehate on Reddit.

Source: MetaFilter

In addition to sites like 4chan and Voat, we also try to use well-moderated sites, like MetaFilter, as distractors or counterexamples of abusive content. MetaFilter requires $5 to establish an account, and is one of the most strictly moderated communities on the Internet. Moderators hide inappropriate material quickly, and reinforce positive norms by making good behavior far more visible than bad [55]. Whenever needed, moderators step in and temporarily suspend an offending user's account.

Source: r/AskHistorians, r/AskScience & r/NeutralPolitics

r/AskHistorians and r/AskScience are communities that are actively moderated, and have well-defined rules regarding user behavior and interactions on the subreddit. These rules are regularly enforced by moderators and exist to ensure that debates on the subreddit do not devolve into personal insults or ad hominem attacks.

r/AskScience urges its users to "Be civil: Remember the human and follow Reddiquette", in its guidelines [45, 46]. r/AskHistorians has a strict "Civility" rule which says, "All users are expected to behave with courtesy and politeness at all times. We will not tolerate racism, sexism, or any other forms of bigotry. This includes Holocaust denialism. Nor will we accept personal insults of any kind." [47]

r/NeutralPolitics is a well-moderated community "dedicated to evenhanded, empirical discussion of political issues." The community urges its users to be courteous in its comment rules,[5] which states that "Name calling, sarcasm, demeaning language, or otherwise being rude or hostile to another user will get your comment removed."

[3] http://redditmetrics.com/r/fatpeoplehate
[4] http://redditmetrics.com/r/CoonTown
[5] https://www.reddit.com/r/NeutralPolitics/wiki/guidelines

Target: MixedBag

We have a research partnership with a large online community who provided data moderated off-site for violating abuse policies. Getting data such as these is typically a major hurdle, as companies fear the blowback that may occur after its release. As per our partnership agreement, we will refer to this target community using a pseudonym: *MixedBag*. The community has on the order of 100M users, and is typical of user-generated content sites: the site has profiles, posts, comments, friends, etc. We obtained comments that were deleted by the site's moderators as abusive, and flagged by users, as part of this partnership.

A notable challenge is that *a priori*, the target and source sites share little in common. For example, MixedBag is a pseudonymous community where conversation is structured into threads of comments, in response to a piece of shared content; 4chan is an image board where anonymous people often post short and unrelated phrases in response.

DATA

We collected data from each of our source communities, as well as data from MixedBag as target data. In our static model, we use the source and target datasets as classic train and test datasets. In the dynamic model, we iteratively allow a model trained on source data to update itself as it sees new batches of target data.

Source data

We collected varying amounts of data from each source community, as it was available:

- **3M posts** from 4chan /b/ and /pol/ boards, spanning 14 months in 2015 and 2016
- **700K posts** from r/fatpeoplehate and r/CoonTown, spanning January to July 2015
- **70K posts** from v/fatpeoplehate and v/[n-word], spanning August 2015 to February 2016
- **2M posts** from MetaFilter, which contains all posts archived on the site, spanning July 1999 to July 2015
- **1.5M posts** from r/AskScience and r/AskHistorians, spanning 2007 to 2015
- **130K posts** from r/NeutralPolitics, spanning 2007 to 2015

We also obtained **3.5M** random comments from MixedBag, which were publicly available at the time of data collection. The comments serve as distractors for building a BoC model: they represent a random sample of the site's publicly visible comments. In total, we collected over **10M** posts to serve as training data using a variety of archives and crawlers. *Note:* our training phase does not give static models access to comments moderated from MixedBag.

Target data

To evaluate our model, we obtained the text in **200,000** *moderated comments* from MixedBag. The dataset contains over 4 years of human-curated data—comments moderated off-site

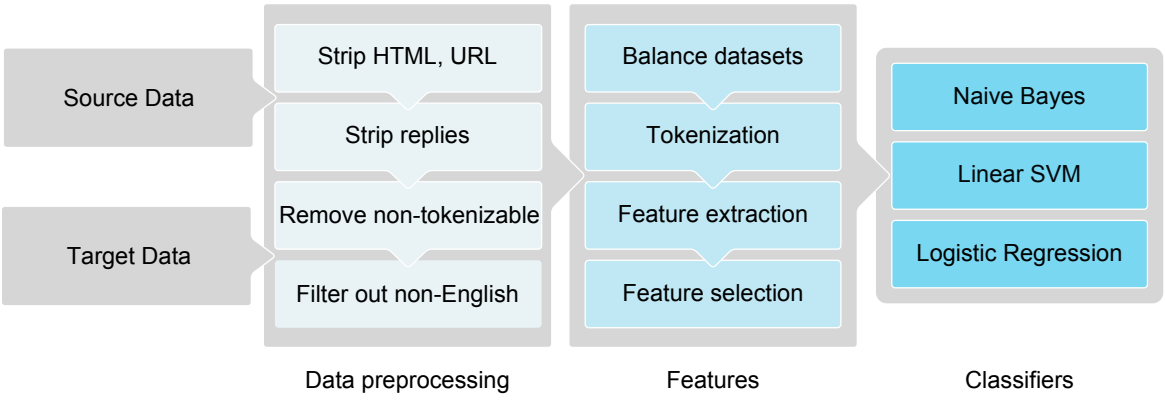

Figure 3. Flowchart depicting the overall *CCS* model-building pipeline. After collecting Bag of Communities and MixedBag data, text undergoes a number of preprocessing steps before acting as input for three different classifiers. Each *CCS* classifier tries to distinguish a source community's posts from a random background cohort of distrators.

by moderators and users for violating abuse policies. As mentioned before, these data were given to us by MixedBag as part of a research partnership. We also obtained the textual content of **200,000** random MixedBag comments, which were still present online during the time of data collection, using the same procedure as in the section above. Note that there is no overlap between (on-site) MixedBag comments used for training and testing.

To provide readers a sense of the types of comments moderated off-site, the following randomly-sampled ones represent typical instances. Readers are forewarned that most are offensive and "not safe for work" (NSFW):

> SO I WILL LOOK YOU'RE FAMILY UP ON WHITEPAGES AND MURDER YOU!!
>
> Lol, go kill yourself
>
> go fuk yourself ugly
>
> YOUR GRANDFATHER IS BURING IN HELL KIKE!!
>
> hehehe! we're gonna have a lot of fun with this! now, lie on your back.
>
> This is full of fail and AIDS!
>
> awwwww what a cute [n-word]

APPLYING BOC TO ABUSIVE BEHAVIOR ONLINE

We used the data collected from our *Bag of Communities* and MixedBag to build and evaluate multiple machine-learning models. In this section, we will discuss the components of our BoC model and steps in the model's pipeline. For reference and overview, Figure 3 visualizes the pipeline for training every internal *CCS* estimator.

Data preprocessing

We began preprocessing the data sets by stripping replies, HTML elements and URLs in the collected comments. Next, we discarded posts that were not tokenizable. These were comments that were either not in Unicode or did not contain any text/tokens. Finally, we performed language detection and discarded comments that were not in English. We used *langdetect* [54], an open-source Java library, for language identification.

Balancing datasets

After the preprocessing steps, we shuffle the datasets and balance them to ensure an equal number of posts from each class. Note that balancing the number of samples from each class likely does not mimic real-world situations. In general, abusive posts are relatively rare. However, balancing across all conditions ensures that we can easily interpret model fits relative to one another. In other words, since the in-domain model will also act on the balanced datasets, balancing will not privilege either approach.

Tokenization & feature extraction

We tokenize comments, and break the text contained in each comment into words. Using these words, we go on to build the vocabulary for all comments in the sample. Each comment is represented as a feature vector of all words and phrases present in the vocabulary (i.e., a *Bag of Words* (BoW) model). The feature values are either the binary-occurrence values (present or not) or the frequency of occurrence. We extract n-grams ($n \in [1, 2, 3]$) from the text and perform vectorization using a Hashing Vectorizer. Hashing Vectorizers create a mapping between tokens and their corresponding feature value (TF-IDF) [43].

Feature selection

We compute the ANOVA F-values for the provided sample, and select the most distinguishing features using the F-value between features and labels. We perform feature selection on features in the *top k* in $[100, 10^3, 10^4, all]$. For example, when *top* $k = 10^3$, only the top $1,000$ BoW features are selected based on their ANOVA F-values.

Classifiers

To build internal *CCS* estimators, we ran classifications tests using three different classifiers: *Multinomial Naive Bayes* (NB), *Linear Support Vector Classification* (LinearSVC), and

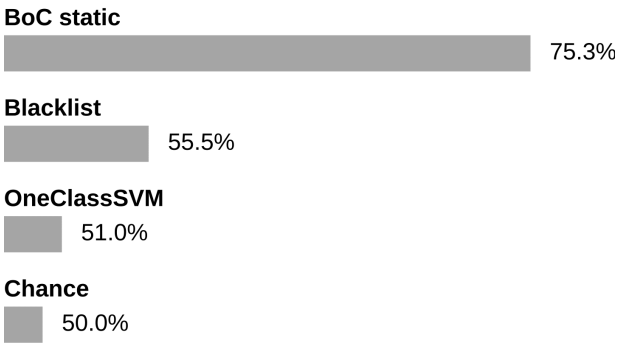

Figure 4. Accuracy values for baselines and the BoC static model. *Chance* refers to a random classifier.

Logistic Regression (Logit). Our BoC models use the output likelihoods from these classifiers as internal estimators.

Parameter search

We ran classification tests on the datasets using different settings to identify the best configuration for our BoC model. We performed a grid search on a held out 5% sample of our data, provided by *GridSearchCV* [43]; it exhaustively generates candidates from a grid of parameter values specified in the grid shown in Table 1. For example, *n-gram range* refers to the range of n-grams extracted, and with *n-gram range* = (1, 3), we extracted uni-grams, bi-grams and tri-grams. These parameter values were used to find the best configuration for our models, and are used in all subsequent phases.

BoC static model

We explore two models in this paper. The first we call the "BoC static model," a model that sees no data from the MixedBag target data at all. This BoC static model trains the underlying *CCS* estimators, but gets no access to test data; therefore, it resembles a pre-trained model that could be deployed "off the shelf," similar to how blacklists are often used in practice today.

Parameter	Values	Best Value
n-gram range	[(1,1), (1,2), (1,3)]	(1,3)
binary	[on, off]	off
lowercase	[on, off]	on
max features	$[2^{22}, 2^{26}]$	2^{26}
tf-idf	[on, off]	on
alpha	[0.1, 0.01]	0.01
feature selection	[on, off]	on
top k	$[100, 10^3, 10^4, all]$	10^4
classifier	NB, LinearSVC, Logit	NB

Table 1. Grid of parameter values used when running classification tests to find the best combination of parameter values for our model. The best values shown for all the parameters, found with a grid search, were used in all classifiers. *max features* refers to the upper limit placed upon the hashing vectorizer.

Model	Precision	Recall	Accuracy
BoC static	77.49%	71.24%	75.27%
In-domain	88.20%	91.66%	89.77%
Only abuse BoC dynamic	95.04%	85.85%	**91.18**%
All BoC dynamic	91.09%	87.93%	90.20%

Table 2. Precision, recall and accuracy for different models. The dynamic (online learning) models were trained on 100,000 test samples.

We built two different baselines to compare with our BoC static model, and arrive at performance (lower) bounds.

BoC static baseline: Blacklist

We first trained a model to classify an input comment as abusive or not based on the presence of blacklisted words. We obtained a list of profane terms used in previous work [58]. Such list-based detection mechanisms are commonly deployed in the wild. This model essentially checks for the presence of at least *threshold* number of blacklisted term(s) in the comment. We tested the model for all values of $threshold \in [1, 2, 3, \ldots]$.

BoC static baseline: OneClassSVM

In the absence of labeled, rare and supposedly different data points, one known approach is treating such points as outliers of a known distribution. In our case, we trained a OneClassSVM [43] to learn the distribution of n-grams for naturally-occurring MixedBag posts, in the hope that abusive posts will deviate from this distribution. The OneClassSVM was trained on just 3.5 million random MixedBag posts, and tested on the target data from MixedBag. The parameter configurations used are shown in Table 1.

BoC dynamic model

In addition to the static model, we also explore a "dynamic" (or online learning) model that iteratively sees more and more target community data to aid prediction. This mimics what an upstart community might face when building its own abuse detection models as new moderator labels come in. The BoC dynamic model uses these data in conjunction with internal *CCS* estimators to make final predictions. That is, it has access to the cross-community similarity scores, $CCS(S_i, T)$, described earlier, which gives the likelihood of a (target) post belonging to a (source) community S_i.

In particular, the dynamic BoC model is provided with similarities to each source community, $CCS(S_i, T)$, which is the *predict_proba*() returned by an internal estimator (NB) trained on source community data. These probability scores are used as features, in addition to textual features learned from the target community data by the final estimator (also NB), which predicts whether a given post is abusive or not. We compare it against a purely in-domain linguistic model with the same parameter setting. Both models—the online, in-domain model and the dynamic BoC model—are trained the same way, on a fractional batch of the target data, and then evaluated on the remaining (unseen) target data.

More formally, at a given iteration where the model sees a fractional batch of size f of the target data, the models are

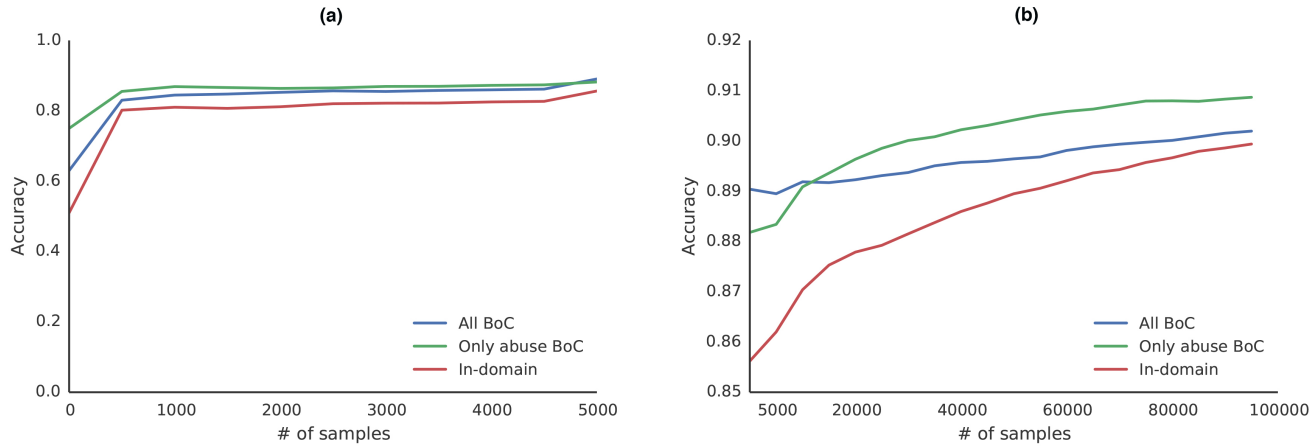

Figure 5. Dynamic model performance when trained only on target community (MB) data and including CCS, BoC features. *In-domain* denotes the plain partial fit model that uses only MB data, *Only abuse BoC* denotes the dynamic model only using communities that are hypothesized abusive, and *All BoC* denotes the dynamic model using all communities in our dataset. Performance of the models when iteratively trained on up to 5,000 target community samples are shown in (a), and the remaining batch sizes in (b). The plots are separated for better resolution, and (b) is scaled up for clarity.

constructed as follows:

$$M_{in} \sim BoW(f \cdot MB)$$

$$M_{dyn} \sim BoW(f \cdot MB) + CCS(S_i, T), \forall S_i \in BoC(all)$$

$$M_{dynabuse} \sim BoW(f \cdot MB) + CCS(S_i, T), \forall S_i \in BoC(abuse)$$

All models build BoW linguistic models of the data to which they have access so far. The *in-domain* model (M_{in} above) is trained only on posts from the target community, and does not see any of the BoC data. The *all BoC dynamic* model (M_{dyn} above), is trained on posts from the target community, in addition to CCS (internal) estimations from all 9 source communities. Whereas the *only abuse BoC dynamic* model ($M_{dynabuse}$ above) uses CCS estimations from only the abusive communities (i.e., 4chan, r/fatpeoplehate, r/CoonTown, v/[n-word], v/fatpeoplehate).

We aimed to observe the growth in accuracy of predictions over time (as more and more moderated posts from the community are available for training the model) and understand when the performance values saturate.

RESULTS

While we ran trials with three different classifiers (see above), Multinomial Naive Bayes (NB) performed best in all conditions. The simplest model, its performance may reflect its limited ability to overfit the training data. Hereafter, we report results for the NB model across conditions. The parameter values used for the best model are available in Table 1.

BoC static model performance

We compared the performance of our best BoC static model with two different baselines. Figure 4 displays the accuracies across models. We observed that the Blacklist gave a best performance of 55% (with threshold 1), while the OneClassSVM achieved an accuracy of 51%. Our BoC static model performed at 75.3% accuracy.

BoC dynamic model performance

The *BoC dynamic* online learning models performed uniformly better than a purely *in-domain* model built only using moderated posts from the target community. The differences in performance of the *in-domain*, *all BoC dynamic*, and *only abuse BoC dynamic* models at various stages of data access are shown in Figure 5. At 0 test samples seen, the in-domain model performed at 51% accuracy (it is equivalent to a single-class classifier used to detect outliers, without any access to moderated posts). The BoC dynamic models outperformed the purely in-domain model even after 100,000 (moderated) test samples were seen. The best performing BoC dynamic model achieved 91.18% accuracy, after seeing 100,000 (moderated) test samples. At all batch sizes measured, the differences are statistically significant.

DISCUSSION

We find that the Granger-causal, CCS-based, *Bag of Communities* models perform well in both static and dynamic settings. The static model likely performs well enough right now that it could be deployed as is with human oversight on a new community; the dynamic model uniformly outperforms purely in-domain classifiers with access to years of curated data. This means that models operating entirely on out-of-domain (4chan, Reddit, Voat and MetaFilter) data can learn significant cross-domain knowledge applicable to a community the model has never seen before. Given that we performed no domain adaptation [2, 12, 13], this result signals deep overlap between, for instance, large-scale preexisting Internet data and comments on another site.

We do not intend to intimate with these results that sites should substitute a BoC model for their existing moderation systems. Rather, this paper presents a promising empirical result about the utility of using preexisting community

data to inform abuse detection. It suggests that gathering data from other communities could be extremely useful. Next, we reflect on our models, discuss some of their error patterns, strategize about selecting source communities, and conclude with reflection on how designers and researchers could use BoC models.

Reflection on models

In post-hoc inspection we observed that our BoC-based model identified a significantly larger variety of *abusive* content than the other models. This is in accordance with the high precision values achieved by the BoC classifiers when classifying abusive content (see Table 2). This derives from the source communities. For instance, the BoC data provides background information not available to the in-domain model, ranging from popular Internet phrases (e.g., "full of fail", "FOR THE LULZ") and terms (e.g., "desu", "nips") to variants of commonly used terms (e.g., "fuk"). Most of these comments were not identified by the in-domain model, as it sees only a handful of such terms in MixedBag posts.

As seen in Table 2, the BoC exhibit better precision-to-recall trade-offs than purely in-domain models. That is, they naturally seem to trade recall for precision more than the in-domain models. In discussions with site operators, this seems to be the way they would prefer the model's error patterns to behave. As many social media companies are owned and operated in the United States, concerns about censorship understandably pervade discussions around moderation [62]. High precision models (i.e., if the model declares it "abusive," then it very likely is, even at the cost of missing more abusive posts on average) would fit well in this context.

Error analysis

Both the in-domain and the BoC models missed a significant fraction of abusive posts. In an error analysis, many of them used character-level substitutions to evade automatic filters (e.g., "f**king", "f**k"), but were identified by human moderators on MixedBag. You could imagine normalization filters that help to uncover substitutions like these [5], a fruitful area for future work improving these models and data pipelines. Run of the mill spam also seemed to evade all models, suggesting that a future enhancement would be to add existing spam filters to data pipeline in Figure 3.

Some moderated posts were sarcastic in nature, and automatic detection of sarcasm is an open research problem [22, 51]. While neither the in-domain models nor the BoC could catch these instances, they were identified by human moderators on MixedBag:

> if i had a dollar for every pixel in this picture, i'd have 50
>
> Oh mai gowd I have never been so enlightened in mai hole laif.
>
> aww your going blind ???

Reflecting the noise of the real world, we also observed the presence of non-abusive posts in our test samples, which were (perhaps wrongly) removed by site moderators. Sometimes, moderators delete entire threads of comments, posted in response to inflammatory or offensive (parent) posts. Examples of likely mislabeled data:

> This is really cool! Superb job < 3
>
> Wonderful job
>
> Its WOW!
>
> Aww! You should nominate him for [award]! See [link] for details, okay?

Best performing model: Only abuse BoC

r/NeutralPolitics, r/AskScience, r/AskHistorians and MetaFilter are all well-moderated communities. We observed that training the *All BoC* model including data from these communities, in addition to the hypothesized abusive communities, increased the number of false positives (i.e., non-abusive posts being misclassified as being abusive). This can be attributed to the fact that typical (onsite) content found in MixedBag is more similar to 4chan, v/[n-word], v/fatpeoplehate, r/fatpeoplehate and r/CoonTown, than the former. In other words, the content found on the former communities are too polite (or well-moderated), and observed to not be representative of the normative behavior in the target community. As a result, the *Only abuse BoC* model achieved the best accuracy in our tests.

Choosing source communities

The choice of 4chan boards, hate-filled subreddits and subverses as source communities required some community-level insight. The intuition that many of these communities perform "bad behavior" motivated our data collection. We have also looked at a variety of well-moderated communities like MetaFilter, r/AskHistorians and r/AskScience, and r/NeutralPolitics as counterexamples.

How do you choose the community data required for BoC? At present, there is some "black magic" involved in collecting the right communities so as to be useful for a given context—not unlike the infrequently discussed black magic surrounding feature engineering in many applied machine learning contexts. For the moment, we believe this will be driven by the problem at hand. Intuition and domain knowledge will likely drive BoC data collection, and more work should be done to explore how to reduce search and collection costs. While community data such as this only needs to be collected once, it does require some investment of time and energy to write crawlers, debug them, etc.

However, it is possible to envision scenarios where many of the Internet's most important and popular communities have been crawled, stored, and used in training *CCS* classifiers. For example, given the encouraging results in the present work, we have recently explored simply building a *CCS* classifier for every subreddit, for all contributions ever posted to that subreddit. Even with a strict threshold on activity level (i.e., only include subreddits above a certain subscriber level, or post level), this would number in the thousands. You can imagine doing something similar among many well-connected communities (in the social network sense [38]) on Twitter. An API could live between an implementing application and all these data sources, essentially generating thousands of *CCS* feature vectors for applications. Scaling up in this manner seems very promising.

Design Implications

We have released our models under an LGPL license. We host the pre-trained internal *CCS* estimators used for both the static and dynamic models we presented here.[6] Even at this early stage in the research, we believe the present BoC models can be used to address the challenge of moderating new, emerging and established communities. As demonstrated in this conceptual and empirical work, the BoC approach may allow communities to deal with a range of common problems, like abusive behavior, faster and with fewer engineering resources.

Communities themselves may choose to operationalize the models we provide in a number of ways. A site might choose to wrap them in a human-in-the-loop system, where moderators review comments triaged by BoC score. Another community with fewer resources may automatically hold comments that score above a certain threshold, requiring users to petition to have them looked at by a human moderator. We hope that the open source models provide the kind of flexibility necessary for site operators to build these and other moderation approaches.

Theoretical Implications

As alluded to in the *Introduction*, the BoC approach may enable new kinds of theoretical advances. Returning to our example of a maximal BoC classifier that returns thousands of similarity features for all of Reddit, these scores place communities in a metric space: they are embedded in a high-dimensional space either closer or farther away from other communities. In this way, this approach might allow a more theoretically-inclined researcher to extract latent clusters of online communities into a taxonomy. An advantage of using semi-automated clustering techniques such as this one is that it would update as the Internet changes, without requiring another full round of researcher effort. We hope to see theoretical work exploring approaches along these lines.

LIMITATIONS & FUTURE WORK

While we find these results encouraging, they raise a number of questions, challenges and issues. Here, we reflect on some of the limitations present in our work, with an eye toward how we and others might build upon it.

Focus on linguistic information. The empirical part of this paper only examines the similarity between words and phrases in two communities. We have left out many pieces of data including temporality, other media such as images, etc. Work exploring and including these data would paint a richer picture of communities, and very possibly aid in prediction tasks. For example, it seems very reasonable to assume that reply chain dynamics (i.e., fast-arriving replies vs. slow-arriving replies) might interact with a post's likelihood of exhibiting abusive properties.

Focus on one target community. Here, we only look at the transference of information between 4chan, Reddit, Voat, MetaFilter, and one community, MixedBag. While we find the results encouraging and surprising, whether BoC-based techniques will work with other communities remains open. While we see it as a positive to have obtained MixedBag data at all, given the challenges facing companies who would release it, we encourage other researchers to explore extending the BoC approach to other target communities. For example, we are currently aware of another research lab looking to extend this work to moderated comments from various subreddits (observable through the open Reddit API).

Human-in-the-loop systems. In real situations, human moderators will still need to supervise any automated triage system, including ones built on BoC. It remains open exactly how well BoC would fare in situations like these, and how best to design it to do so. We see this as very profitable future line of work for human-centered computing researchers, as it will certainly involve talking with and understanding the work of professional and amateur community moderators.

CONCLUSION

We presented a new method for transferring one community's data to another—the *Bag of Communities* approach—and apply it to a key challenge faced by communities and moderators. To the best of our knowledge, the BoC representation is novel within social computing and applied machine learning more generally. Further, we presented an existence proof of Bag of Communities: building a classifier for identifying abusive content using data from 4chan, Reddit, Voat and MetaFilter. We hope that other designers find BoC useful in predicting a variety of important online phenomena; we hope researchers can make use of it to examine other communities with the Bag of Communities representation.

ACKNOWLEDGMENTS

We thank the social computing group at Georgia Tech, the Coral Project, Nate Matias and the Media Lab workshop for their valuable inputs that improved this work. This work is supported by NSF grant IIS-1553376.

REFERENCES

1. Ricardo Baeza-Yates, Berthier Ribeiro-Neto, and others. 1999. *Modern information retrieval*. Vol. 463. ACM press New York.
2. Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning* 79, 1-2 (2010), 151–175.
3. Michael S Bernstein, Andrés Monroy-Hernández, Drew Harry, Paul André, Katrina Panovich, and Gregory G Vargas. 2011. 4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online Community.. In *ICWSM*. 50–57.
4. Sam Biddle. August, 5, 2015. Reddit (Finally) Bans CoonTown. http://gawker.com/reddit-finally-bans-coontown-1722332877, *Gawker* (August, 5, 2015).
5. Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In

[6] https://bitbucket.org/ceshwar/bag-of-communities.git

Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 562–570.

6. Brooks Buffington. April 4, 2015. Personal communication. (April 4, 2015).
7. Catherine Buni and Soraya Chemaly. 2016. The Secret Rules of the Internet, Apr. 2016. *http://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech* (2016).
8. Nan Cao, Conglei Shi, Sabrina Lin, Jie Lu, Yu-Ru Lin, and Ching-Yung Lin. 2016. TargetVue: Visual Analysis of Anomalous User Behaviors in Online Communication Systems. *Visualization and Computer Graphics, IEEE Transactions on* 22, 1 (2016), 280–289.
9. Stevie Chancellor, Zhiyuan Jerry Lin, and Munmun De Choudhury. 2016. This Post Will Just Get Taken Down: Characterizing Removed Pro-Eating Disorder Social Media Content. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 1157–1162.
10. Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial Behavior in Online Discussion Communities. In *Ninth International AAAI Conference on Web and Social Media*.
11. Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proc. ACL'13*.
12. Hal Daumé III. 2009. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815* (2009).
13. Hal Daume III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research* (2006), 101–126.
14. Nicholas Diakopoulos. Apr. 2015a. Picking the NYT Picks: Editorial criteria and automation. (Apr. 2015).
15. Nicholas A Diakopoulos. 2015b. The Editor's Eye: Curation and Comment Relevance on the New York Times. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1153–1157.
16. Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of Textual Cyberbullying.. In *The Social Mobile Web*. 11–17.
17. Roger Dingledine, Nick Mathewson, and Paul Syverson. 2004. *Tor: The second-generation onion router*. Technical Report. DTIC Document.
18. Bruce Drake. 2014. The darkest side of online harassment: Menacing behavior. *Pew Research Center, http://www.pewresearch.org/fact-tank/2015/06/01/the-darkest-side-of-online-harassment-menacing-behavior/* (2014).
19. Maeve Duggan. 2014. Online harassment: Summary of findings. *Pew Research Center, http://www. pewinternet. org/2014/10/22/online-harassment/(accessed 02 June 2015)* (2014).
20. Randy Farmer and Bryce Glass. 2010. *Building web reputation systems*. " O'Reilly Media, Inc.". 243–276 pages.
21. Eric Gilbert. 2013. Widespread underprovision on reddit. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 803–808.
22. Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in Twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, 581–586.
23. Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* (1969), 424–438.
24. Jiawei Han, Micheline Kamber, and Jian Pei. 2011. *Data mining: concepts and techniques*. Elsevier.
25. Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. 2002. Searching for safety online: Managing "trolling" in a feminist forum. *The Information Society* 18, 5 (2002), 371–384.
26. Lauren Hockenson. July, 9, 2015. What is Voat, the site Reddit users are flocking to? http://thenextweb.com/insider/2015/07/09/what-is-voat-the-site-reddit-users-are-flocking-to/, (July, 9, 2015).
27. Ruogu Kang, Laura Dabbish, and Katherine Sutton. 2016. Strangers on Your Phone: Why People Use Anonymous Communication Applications. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 359–370.
28. Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design. MIT Press, Cambridge, MA* (2012), 125–178.
29. Amy Jo Kim. 2000. *Community building on the web: Secret strategies for successful online communities*. Addison-Wesley Longman Publishing Co., Inc.
30. Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 543–550.
31. Lawrence Lessig. 1999. *Code and other laws of cyberspace*. Vol. 3. Basic books New York.

32. Jing-Kai Lou, Kuan-Ta Chen, and Chin-Laung Lei. 2009. A collusion-resistant automation scheme for social moderation systems. In *Consumer Communications and Networking Conference, 2009. CCNC 2009. 6th IEEE*. IEEE, 1–5.
33. Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, and others. 2008. *Introduction to information retrieval*. Vol. 1. Cambridge university press Cambridge.
34. Bamshad Mobasher, Robin Burke, Runa Bhaumik, and Chad Williams. 2007. Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness. *ACM Transactions on Internet Technology (TOIT)* 7, 4 (2007), 23.
35. Justin Wm. Moyer. July, 17, 2015. Coontown': A noxious, racist corner of Reddit survives recent purge. https://www.washingtonpost.com/news/morning-mix/wp/2015/07/17/coontown-a-noxious-racist-corner-of-reddit-survives-recent-purge/, *The Washington Post* (July, 17, 2015).
36. Lev Muchnik, Sinan Aral, and Sean J Taylor. 2013. Social influence bias: A randomized experiment. *Science* 341, 6146 (2013), 647–651.
37. Edward Newell, David Jurgens, Haji Mohammad Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. 2016. User Migration in Online Social Networks: A Case Study on Reddit During a Period of Community Unrest. In *Tenth International AAAI Conference on Web and Social Media*.
38. Mark EJ Newman. 2006. Modularity and community structure in networks. *Proceedings of the national academy of sciences* 103, 23 (2006), 8577–8582.
39. Fox News. Apr. 2009. 4Chan: The Rude, Raunchy Underbelly of the Internet. http://www.foxnews.com/story/2009/04/08/4chan-rude-raunchy-underbelly-internet.html. (Apr. 2009).
40. Elinor Ostrom. 2015. *Governing the commons*. Cambridge university press.
41. Ellen Pao. July 16, 2015. Former Reddit CEO Ellen Pao: The trolls are winning the battle for the Internet. http://wapo.st/1HJM82l, (July 16, 2015).
42. Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting comment moderators in identifying high quality online news comments. In *Proc. Conference on Human Factors in Computing Systems (CHI)*.
43. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and others. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* 12 (2011), 2825–2830.
44. Jenny Preece and Diane Maloney-Krichmar. 2003. Online communities: focusing on sociability and usability. *Handbook of human-computer interaction* (2003), 596–620.
45. Reddit. 2014. Remember the human. https://www.reddit.com/r/blog/comments/1ytp7q/remember_the_human/. (2014).
46. Reddit. 2015a. Reddiquette. https://www.reddit.com/wiki/reddiquette. (2015).
47. Reddit. 2015c. Subreddit rules. https://www.reddit.com/r/AskHistorians/wiki/rules. (2015).
48. Reddit. June, 10, 2015b. Removing harassing subreddits (self announcement). https://www.reddit.com/r/announcements/comments/39bpam/removing_harassing_subreddits/, (June, 10, 2015).
49. Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. 2000. Reputation systems. *Commun. ACM* 43, 12 (2000), 45–48.
50. Paul Resnick and Richard Zeckhauser. 2002. Trust among strangers in internet transactions: Empirical analysis of ebays reputation system. *The Economics of the Internet and E-commerce* 11, 2 (2002), 23–25.
51. Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as Contrast between a Positive Sentiment and Negative Situation.. In *EMNLP*, Vol. 13. 704–714.
52. Adi Robertson. June, 10, 2015. Reddit bans 'Fat People Hat' and other subreddits under new harassment rules. http://www.theverge.com/2015/6/10/8761763/reddit-harassment-ban-fat-people-hate-subreddit, *The Verge* (June, 10, 2015).
53. P. Shuman. Jul 2007. Fox 11 investigates: anonymous. https://www.youtube.com/watch?v=DNO6G4ApJQY. (Jul 2007).
54. N. Shuyo. 2010. Language detection library for java. (2010).
55. Leiser Silva, Lakshmi Goel, and Elham Mousavidin. 2009. Exploring the dynamics of blog communities: the case of MetaFilter. *Information Systems Journal* 19, 1 (2009), 55–81.
56. Christine B Smith, Margaret L McLaughlin, and Kerry K Osborne. 1997. Conduct control on Usenet. *Journal of Computer-Mediated Communication* 2, 4 (1997), 0–0.
57. Sara Sood, Judd Antin, and Elizabeth Churchill. 2012a. Profanity use in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1481–1490.
58. Sara Owsley Sood, Judd Antin, and Elizabeth F Churchill. 2012b. Using Crowdsourcing to Improve Profanity Detection.. In *AAAI Spring Symposium: Wisdom of the Crowd*.

59. Sara Owsley Sood, Elizabeth F Churchill, and Judd Antin. 2012c. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology* 63, 2 (2012), 270–285.

60. R. Sorgatz. 2009. Macroanonymous is the new microfamous. http://fimoculous.com/archive/post-5738.cfm. (2009).

61. Abhay Sukumaran, Stephanie Vezich, Melanie McHugh, and Clifford Nass. 2011. Normative influences on thoughtful online participation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3401–3410.

62. Nabiha Syed and Ben Smith. Jun. 2015. A First Amendment For Social Platforms. https://medium.com/@BuzzFeed/a-first-amendment-for-social-platforms-202c0eab7054. (Jun. 2015).

63. Nitasha Tiku and Casey Newton. February 4, 2015. Twitter CEO:We suck at dealing with abuse.. http://www.theverge.com/2015/2/4/7982099/twitter-ceo-sent-memo-taking-personal-responsibility-for-the, *The Verge* (February 4, 2015).

64. Ruth L Williams and Joseph Cothrel. 2000. Four smart ways to run online communities. *MIT Sloan Management Review* 41, 4 (2000), 81.

65. Jun-Ming Xu, Benjamin Burchfiel, Xiaojin Zhu, and Amy Bellmore. 2013. An Examination of Regret in Bullying Tweets.. In *HLT-NAACL*. 697–702.