

Profanity Use in Online Communities

Sara Owsley Sood
Computer Science Dept., Pomona College
185 East Sixth Street
Claremont, CA 91711
sara@cs.pomona.edu

Judd Antin, Elizabeth F. Churchill
Yahoo! Research
4301 Great America Parkway
Santa Clara, CA 95054
<jantin, echu>@yahoo-inc.com

ABSTRACT

As user-generated Web content increases, the amount of inappropriate and/or objectionable content also grows. Several scholarly communities are addressing how to detect and manage such content: research in computer vision focuses on detection of inappropriate images, natural language processing technology has advanced to recognize insults. However, profanity detection systems remain flawed. Current list-based profanity detection systems have two limitations. First, they are easy to circumvent and easily become stale—that is, they cannot adapt to misspellings, abbreviations, and the fast pace of profane slang evolution. Secondly, they offer a one-size fits all solution; they typically do not accommodate domain, community and context specific needs. However, social settings have their own normative behaviors—what is deemed acceptable in one community may not be in another. In this paper, through analysis of comments from a social news site, we provide evidence that current systems are performing poorly and evaluate the cases on which they fail. We then address community differences regarding creation/tolerance of profanity and suggest a shift to more contextually nuanced profanity detection systems.

Author Keywords

Online communities, comment threads, user-generated content, negativity, community management, profanity.

ACM Classification Keywords

H.5.3. Information interfaces and presentation: Group and Organization Interfaces.

General Terms

Design, Experimentation, Human Factors.

INTRODUCTION

Online communities are often plagued with negative content – user-generated content that is negative in tone, hurtful in intent, mean, profane, and/or insulting. Negative content can be problematic for sites wanting to expand their

user base, engage existing users, and foster a positive and collaborative community. Social contracts and normative behaviors, however, are unique to specific socio-technical systems. What is considered inappropriate in a given context is both site and community specific. On many sites, community managers are primarily responsible for the task of removing inappropriate content. However, the flood of user-generated content on many sites quickly overwhelms community managers' ability to effectively manage it.

The detection of negative content of malicious intent (personal attacks and insults) in forums and comment streams is a challenging and nuanced problem [20]. Recent work in machine learning and natural language processing has approached this task with varying degrees of success: with maximal f-measures of 0.298 for detection of harassment on Slashdot and 0.313 on MySpace from one study [24] and a maximal f-measure of 0.5038 for detection of personal insults on another [19].¹ Given the recent attention to the complex and sometimes grave consequences of cyberbullying [1], the ability to recognize and potentially mitigate profanity and other forms of harmful negativity in user-generated content is more important than ever [3,14].

Compared to the challenges of detecting malicious content or spam, detection and removal of profanity is often thought to be an easier task. Most current approaches to profanity detection check new content against large lists of profane terms. However, these systems are flawed in at least two major ways. First, static term-lists quickly lose currency and are relatively easy to circumvent. Users often disguise or partially censor profanity by replacing one or more letters with punctuation marks (e.g., “@ss,” “%@#s”). Misspellings, both intentional or unintentional (e.g. “biatch”, or “shiiiiit”) and the use of slang words (e.g. “assbite”) that evolve quickly and often have local variations also challenge profanity lists to be more thorough and adaptive than they can reasonably be. Thus, these systems face issues of recall; they are unable to catch most cases of profanity. Secondly, list-based approaches to profanity detection are a one-size-fits-all solution that does not take into account differences in community norms and

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CHI '12, May 5–10, 2012, Austin, Texas, USA.

Copyright 2012 ACM 978-1-4503-1015-4/12/05...\$10.00.

¹ F-measure is a measure of accuracy, specifically the harmonic mean of precision and recall.

practices. After all, what constitutes profanity differs greatly based on the specific community and topic at hand. For example, in a forum about dog breeding, “bitch” is a term of art that refers to a female dog, while in many other contexts it is a profane term. Furthermore, sites for children have a drastically different tolerance of profanity than those for adults.

In this paper we make three primary contributions to research on profanity detection. First, we address the state of current list-based profanity detection systems. Do these systems suffice? In what cases do they fail? Secondly, under the assumption that a major oversight in these systems is a lack of tailoring for specific communities, we examine how profanity use differs between communities. Is profanity used more or less in some communities? Do certain communities use profanity in different ways? And finally, we explore the social context of profanity use in different topical communities. How might specific communities receive profanity differently? In the sections that follow, we examine these questions through analysis of a data set of comments from a social news site.

BACKGROUND

As more and more of the web has grown to include user-generated content, the detection and management of inappropriate or objectionable content has become an important task for web sites. One common technique is social moderation, in which users themselves undertake the task of identifying and flagging of profane or inappropriate responses. However these systems have been only moderately successful, and suffer from potential collusion - flagging can be used to indicate disagreement or dislike of a post that is not otherwise inappropriate or profane [15]. Instead of relying on social moderation, recent proposals have been made to automate the detection of inappropriate or abusive content.

Research in computer vision has given much attention to the related issue of detecting inappropriate videos and images. Advances in this space have largely included systems that detect “too much skin” in images and videos [4,11,23]. Other systems utilize textual metadata [8,9], while some combine the two; one such system, *WebGuard* has reached 97.4% accuracy in detecting pornographic web sites [5].

While many would argue that textual analysis is more tractable than visual content analysis, this may be in part because of a general misunderstanding about how difficult the problem of profanity detection is in real-world contexts. Furthermore, text has a visual element that is socially understood. Expressive forms such as emoticons and “ASCII art” use visual properties of text, punctuation marks and symbols to mimic lexical units and thus convey meaning, denote profanity and circumvent automatic filters. Such visual-for-textual substitution is best illustrated through examples such as the use of “@” in “@ss”..

Because of these misunderstandings, perhaps, comparatively little research has focused on detecting inappropriate text in user-generated content systems. As mentioned above, two groups have built systems to detect insults and harassment in online forums [19,24] and another has focused on cyberbullying of teens [3], but even fewer have addressed the identification of profanity. Yoon et al. built a system to detect newly coined profanities in Korean, in an attempt to improve upon the failure of list-based systems to evolve along with profane language [25]. Due to the target audience being children, some have analyzed the content of video game web sites and video games themselves to verify that presented content meets ratings standards [6,7]. However, work in this area does not generally strive for automated analysis.

Advancing our ability to detect and remove profanity could have several significant, positive social consequences. The growth of collaborative information products such as Wikipedia, Yahoo! Answers, and Stack Overflow rely on the provision of interaction environments that are supportive, productive, and meet the specific needs of their user communities. Open-source software projects also rely on email lists and forums to support the necessary community building, coordination, and decision-making processes.

No automated system, by itself, can appropriately filter and manage ongoing discourse and interaction so that it meets the needs of a particular topic, domain, or user community. Indeed, research has illustrated the important role of established community members for implicitly and explicitly communicating language norms to new members [16]. The enforcement of these norms is often ad hoc, however. In large systems, the sheer volume of content means ad hoc strategies often leave a large amount of profane or inappropriate user-generated content undetected. The existence of such content can actually fight against the positive influence of community managers and long-time participants by setting a bad precedent that communicates to new users that profanity and other negative content is acceptable [21]. Automated systems that help community managers, moderators, and administrators to manage the flood of user-generated content in these environments could help to promote more productive large-scale collaboration and thus more valuable information products.

DATA SET

Social news sites (e.g., reddit.com, digg.com) typically allow users to post links to stories of interest, vote on contributed stories, and most important to the present study, comment on stories and respond to others’ comments. Our data set is the complete set of user-contributed comments over a three-month period (March to May 2010) to Yahoo! Buzz, a social news commenting site that is no longer active. Our data set contains 1,655,131 comments distributed among 168,973 distinct threads. In addition to the comment itself, our data set contains meta information

about each comment including the time posted and which news story the comment is in response to; this information can be combined to reproduce the comment thread. We also have information about each news story including its country of origin, language, and category (i.e., politics, entertainment). More information about this data set, including the distributions of comment lengths, comments per user and comments per thread can be found in [19].

Coding

In order to generate a data set describing the presence of profanity, insults, and the objects of the insults, we employed Amazon Mechanical Turk (MTurk). MTurk is an online labor market in which requesters post jobs that can be easily decomposed into a large number of small tasks. MTurk workers (“Turkers”) are presented with a short description of available tasks, and then choose which tasks to complete. Individual tasks typically take between 5 and 20 seconds to complete and workers are generally paid about 5 cents for each task.

Recent studies have suggested that using MTurk for similar content analysis tasks can be both faster and more economical than using dedicated trained raters [22]. Furthermore, several studies have illustrated that combining the work of multiple Turkers on the same task can produce high quality content analysis results, even when some coders do not agree (i.e. the coded data is “noisy”) [2,17].

We selected a random sample of 6500 comments spanning all categories. Comments that were likely to be too short to meaningfully interpret or too long to quickly process were not sampled: we restricted sampling to the 2nd and 3rd quartiles for overall comment length (between 73 and 324 characters long).

Each worker was shown one comment at a time. For each comment, they were asked to answer the following questions:

*“Does this message contain any content you would describe as ‘profanity?’ (including profanity that is disguised or modified such as @ss, s***, and biatch) (Yes/No)*

Thinking about the intent of the comment's author, does the message contain what you would describe as a direct “insult?” (Yes/No)

In your opinion, is the insult directed at the author of a previous comment? (Yes/No/Unsure)

Finally, beyond the requirement of a consensus threshold from multiple coders, we also employed a ‘gold data’ model to improve label quality. Gold data were a set of comments for which the ‘correct’ labels (answers to the above three questions) were designated by the researchers prior to the labeling task. If a Turker mislabeled one of the gold comments, he was shown a short explanation for the correct answer. In this way the gold data functioned as a lightweight training program. In addition, if any Turker incorrectly labeled too many of the gold comments, their

labels were removed from the data set and they were barred from labeling any further comments. All three authors independently judged the ‘correct’ labels for gold comments. For most gold comments the authors agreed on an answer which was likely to be self-evident. For example, the following comment was judged to contain profanity but no insult:

“Now they'll just release them so they can do the same thing tomorrow. Mine the Effen Border. Use the Natl. Guard to patrol our borders with extreme prejudice. Fences don't work. They'll tunnel under them or use ladders. Show the Drug Cartels that we mean business and this sh!t will cease.”

In other cases, however, the primary purpose of the gold comment was to draw attention to the desired aspect of the comment. For example, in the following comment it is, arguably, not possible to conclusively determine who is being insulted:

“Hot off the presses...straight fom their leader oweduma's mouth. The state of our economy deserves attention. This guy must live in a bunker. Wake up you liberal loser! The economy sucks.”

Over the course of approximately 5 days, 221 MTurk workers provided 25,965 judgments on 6500 comments. Following the model suggested by Sheng and colleagues [17], we employed multiple coders for each item. As a result, each item was rated by a minimum of three raters. We adopted a simple consensus model on the labels. To ensure labeling accuracy, our final profanity labeled data set only includes those comments for which at least 66% of labelers agreed on the profanity label. Similarly, our final insult and insult object labeled data sets only include those comments for which at least 66% of labelers agreed on the insult or insult object label. This method resulted in a different N depending on the focal phenomena (profanity, insult, or insult object). For example, one hundred and forty-six comments (2.2%) were dropped from the final profanity labeled data set because raters did not reach consensus on the profanity label.

DO CURRENT PROFANITY DETECTION SYSTEMS SUFFICE?

The standard approach to profanity detection in online communities is to censor or filter text based on lists of profane terms. When user-generated text contains listed words, those words or the entire contribution may be flagged for review or automatically removed. Some profanity lists are shared between multiple sites, and administrators contribute additional terms as they become prevalent or problematic. In order to test the efficacy of this approach, we downloaded a shared profanity list from the site phorum.org and built a simple system that flags a comment as profane if it contains any of the words on the phorum.org list.²

² At the time of our analysis (July 1, 2011), the phorum.org list contained 120 profane terms.

System	Precision	Recall	F1	Accuracy
noswearing.com w/ stemming	0.528	0.402	0.457	0.907
noswearing.com or phorum.org w/ stemming	0.490	0.412	0.448	0.902
noswearing.com or phorum.org	0.516	0.390	0.444	0.906
noswearing.com	0.563	0.367	0.444	0.911
phorum.org w/ stemming	0.631	0.231	0.338	0.913
phorum.org	0.636	0.196	0.300	0.912
random	0.096	0.501	0.161	0.498
weighted random	0.106	0.109	0.108	0.825

Table 1: An evaluation of list-based profanity detection systems.

As noted above, list-based systems often suffer (by identification/“recall” measures) as profane language evolves over time with slang and Internet abbreviations. As such, we downloaded a second list of profane terms from noswearing.com. This site hosts a list of community contributed profane terms. This list evolves over time with user contributions and is larger than the phorum.org list.³ While both lists contain traditional profane terms, they also contain inappropriate terms such as racial slurs and vulgarities.

In another attempt to improve recall, we employ a stemmer. Beyond simply looking for the presence of a word on a profanity list, the stemmer allows the system to see if any words in a comment have a shared stem with any word on a profanity list. To evaluate the efficacy of list-based methods we built several systems that employed the two lists and stemming in various combinations. For each system, we average its performance over 5 trials of 10-fold cross validation on our 6500 profanity-labeled comments from Yahoo! Buzz. While the data set as a whole contains 6500 comments, 6354 meet the 66% labeling consensus across the MTurk labelers for the profanity label. Of those 6354, 595 (9.4% of the corpus) are positive cases, meaning that they contain profanity. All systems are evaluated based on their precision (a measure of false positives), recall (a measure of false negatives), f-measure (f1 – the harmonic mean of precision and recall) and accuracy—this is the standard array of evaluation metrics for systems of this type [10,18,24].

The performances of all systems are summarized in Table 1, sorted in descending order of f-measure. The *random*

³ As of July 1, 2011, the noswearing.org list contained 341 terms.

ass	asses	bullsh!t
s***	pussy	sob
f*****	cr	bitch
azz	stfu	schit
bastard	dumbass	sh
sh!t	\$hit	r**h***
a*****	f***	c***
nr	m*****f*****	phag!
b*****	f*cking	f'ing
sh*t	@ss	goddamn
a\$\$	pussies	bull\$hit

Table 2: Top words distinguishing profane from non-profane comments in the data set.

system is included as a baseline system; it randomly labels comments as profane or not. Similarly, the *weighted random* system labels comments randomly, weighted by the distribution of profane/non-profane comments in the training set. The performances of these systems are included for comparison purposes, though they of course approach the theoretical random baselines. The remaining systems are list-based approaches, based on the lists gathered from phorum.org and from noswearing.com. In an attempt to reach higher recall, we created additional systems that marked a term as profane if it appeared in either one of the two lists. Finally, in some systems we combined word lists with stemming.

While a peak accuracy of 0.913 seems promising, recall that 9.4% of the comments in our corpus contain profane terms. For this testing data, if one built a system that, given a comment, always returned a negative classification (indicating that the comment does not contain profanity), it would have an accuracy of 0.906 as 90.6% of the testing corpus is comments that do not contain profanity. Therefore, f1, precision, and recall are much more descriptive evaluation metrics. As seen in table 1, peak performance of the list-based approaches is reached using the profane terms list from *no swearing.com* combined with a stemming system. This system detected 40.2% of the profanity cases at 52.8% precision. Based on our results, we must conclude that even the best of these list and stemmer based systems would not perform well at detecting and removing profanity in user-generated comments.

WHY DO LIST-BASED APPROACHES PERFORM SO POORLY?

As we have already discussed, list-based approaches perform poorly because of three primary factors: misspellings (both intentional and not), the context-specific nature of profanity, and quickly shifting systems of discourse that make it hard to maintain thorough and accurate lists. To exemplify these problems, we analyzed the words that most commonly distinguish profane from

non-profane comments in our MTurk profanity labeled data set. The top words, seen in Table 2, were sorted in descending order by x , calculated as follows:

$$x = \frac{(posFeatureCt/TotalPosFeatures)}{(negFeatureCt/TotalNegFeatures)}$$

where $posFeatureCt$ is the number of times the word occurred in positive (profane) comments, $negFeatureCt$ is the corresponding value for negative (non-profane) comments, $TotalPosFeatures$ is the sum of all feature counts across all words in the positive comments, and $TotalNegFeatures$ is the corresponding value for the negative comments. The latter values are included to adjust for differences in the profane and non-profane corpora sizes.

Of the thirty-three words in this table, notice that only eight are standard spellings that would be found in listings of profane terms (i.e., ‘ass,’ ‘bastard,’ ‘asses,’ ‘pussies,’ ‘pussy,’ ‘dumbass,’ ‘goddamn’ and ‘bitch’). There are six instances of slang abbreviations for profane terms – ‘sob,’ ‘nr,’ ‘cr,’ ‘sh,’ ‘fing,’ and ‘stfu.’ The remaining nineteen terms are disguised or author censored profanity (e.g., ‘bullsh!t,’ ‘azz,’ ‘f*****’). Thus, a list-based profanity detection system, such as the ones evaluated in the previous section, would fail to catch twenty-five of the top thirty-three profane terms (76%) used in our data set. While these words could, of course, be added to a profanity list for future detection via a list-based system, there are countless further ways to disguise or censor words. This makes representing them in a list a significant challenge.

As further evidence of how widespread the particular problem of disguised or partially censored profanity is, we analyze use of one specific character, the @ symbol. The popularity of Twitter and other social media have resulted in adaptations and specializations of language for online communication [12]. Just as text messaging has an established shared dictionary of acronyms, social media share some community established abbreviations that allow users to pack more content into short messages. One such abbreviation is the @ symbol. When a user writes @rick, they are directing their message to ‘rick’, but in a public medium. The @ symbol provides a short and easy mechanism for directing public comments towards specific individuals, but also helps to bridge the gap between directed and undirected interaction in computer-mediated communication. For example: “@xeeliz Check this out! http://yhoo.it/rq1y2u #NBAFinals.” Two more example tweets are shown in Figure 1. The top tweet from edchi, shows a use of the #. The bottom tweet from kevinmarks includes a use of the @ symbol, indicating that this tweet is directly addressing the user feliciaday.

To study how the @ symbol is used within our complete data set of 1.65 million comments, we looked at all

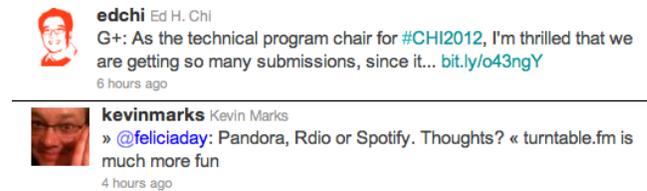


Figure 1: Examples of two tweets, illustrating the use of #, @ and http://bit.ly.

Context	Count of Occurrences of '@'	% of @ usage	% of full data set in this context
Email address	1,112	10%	0.067%
Web address	2,195	19.8%	0.133%
Profanity	4,429	39.9%	0.268%
Conversational	2,764	24.9%	0.167%
Other	592	5.3%	0.036%
Total	11,092	100%	0.67%

Table 3: Analysis of '@' symbol usage within the data set.

comments that contained an instance of the symbol. We found that usage of the @ symbol is somewhat common, however, as you might imagine, not all uses of the @ symbol were in the conversational manner presented above. Some comments contain email addresses (e.g. “john@somecompany.com”) or direct readers to a website (e.g. “@ www.cnn.com”). We also found that comments often use the @ symbol to disguise (e.g. “@ss”) or censor (e.g. “@%#\$”) profanity – one of the very problems that plagues the profanity detection systems described above.

To explore the multiple uses of the @ symbol we built a rule based system using regular expressions. Classifying @ usage as within email or web addresses is easily accomplished with regular expressions, however, automatically determining that @ss is profanity while @john is conversational is a more difficult task. We employ a corpus of profane terms (the lists from phorum.org and noswearing.com), along with a tool to calculate the Levenshtein edit distance between two terms [13]. This calculation adds the number of letter insertions, deletions and changes to transform one word into another. When a term contains the @ symbol, in order to determine if it is profanity, we check to see if the Levenshtein edit distance between the term and any known profane term is equal to the number of punctuation marks present in the term. For example @ss has one punctuation mark (@) and has an edit distance of one from the profane term ‘ass.’ \$%#@ has four punctuation marks and has an edit distance of four from any four letter profane term. Using this approach, we have a very high precision tool that takes a term containing the @ symbol and determines if it is a profane term (either disguised or censored). The recall of this tool is only as good as our list of profane terms.

Category	Profanity		Insult		Directed Insult	
	Occurrence (%)	χ^2	Occurrence (%)	χ^2	Occurrence (%)	χ^2
Overall	9.28	--	20.73	--	10.87	--
Politics	10.70	6.73†	26.80	72.92***	14.30	32.73***
News	9.90	1.83	21.60	1.13	11.40	2.39
Business	9.70	1.29	16.70	11.35**	9.50	2.01
Entertainment	9.30	0.00	18.70	2.98	9.10	3.67
Health	9.00	0.64	14.10	4.05	4.80	4.95
Lifestyle	7.90	0.51	10.70	9.73**	1.70	7.96*
World	7.70	0.01	19.00	1.94	9.10	0.75
Science	6.70	1.98	14.60	6.32	9.90	0.71
Travel	5.60	0.23	18.80	0.00	6.70	0.20
Sports	5.20	6.50	14.70	7.07	3.80	12.90**

Table 4: The distribution of comments containing profanity within topical story domains. Reported χ^2 values are the results of the comparison of profanity, insult, and directed insult frequency within a given category to the frequency across all other categories. Throughout this paper, reported significance values are Bonferroni adjusted where there are multiple comparisons.

*** $p \leq .001$, ** $p \leq .01$, * $p \leq .05$, † $p \leq .1$

Using this tool, we labeled all uses of the '@' symbol in our corpus with 'email address,' 'web address,' 'profanity,' 'conversational,' or 'other' (for instances that were not profanity but also did not appear to take the form of a conversational usage of '@'). The results can be seen in Table 3. First, note that only 0.67% (11,092) of all comments in the data set (1,655,131) contain an '@' symbol. Within this set, 39.9% of '@' usage was within the context of a censored or disguised profane term, while only 24.9% of '@' usages appear in a conversational context.

Nearly 40% of all occurrences of the @ symbol came in the form of disguised or author-censored profanity. The @ symbol is just one of many punctuation marks that could be used to disguise profanity. Moreover, the @ symbol is one that is thought to be commonly used in social media in a conversational manner, yet an astonishing 40% of its uses within our data come in the form of disguised profanity. This is likely to be a conservative estimate, as it is known that list-based measures suffer in recall, as shown in the previous section.

HOW FREQUENTLY IS PROFANITY USED?

In addition to facing issues of recall, list-based approaches are a one-size fits all approach that do not take into account how profanity is used within different domains, contexts, and communities. Through our MTurk labeled data set, we explore the use of profanity in comments on news stories in order to understand more about the frequency and context of profanity use and how it is received.

First, we examine the prevalence of profanity in different topical domains. Dividing our 6500 labeled comments by domain of the article they are in reference to, we see that comments on political stories contain more profanity, more

insults, and more directed insults (directed at authors of previous comments) than in any other domain.

Table 4 shows the distribution of profanity, insults and directed insults in comments within the different domains. To avoid the possibility of Type I error we applied the conservative Bonferroni adjustment to all significance values reported. For clarity, the first value in Table 4 can be read as 10.7% of political comments contain profanity. As previously discussed, the N differs between profanity, insults, and directed insults because, for each we use only items for which coders reached consensus. For profanity, N is 4409, for insults N is 4177 and for directed insults N is 3974. Each comment in the table 4 analyses was labeled with one of the 10 categories shown. Profanity usage in political comments is significantly more common than in other comments. Political comments contain significantly more insults and directed insults than in other domains. On the other end of the spectrum, the lifestyle comments contain significantly fewer insults and directed insults than other domains. The business domain also held significantly fewer insults, while the sports domain had significantly fewer directed insults. As expected, from these data we can conclude that different domains of news story incite varying amounts of profane language and use of insults (general insults as well as those directed at other community members).

IN WHAT CONTEXT IS PROFANITY USED?

Given that our data set contains insult and insult object labels in addition to profanity labels, these labels will be utilized as a measure of context. To further understand how profanity is used within our data set, we investigate the co-occurrence of the 'comment contains profanity' label with 'comment contains insult' and 'comment contains directed insult' labels. If a comment is labeled as both profane and

	% w/ insult	% w/ directed insult
profane	58.66	39.49
non-profane	16.19	8.15

Table 5: Of all profane and non-profane comments, this table presents the breakdown of those that contain an insult or a directed insult.

	% w/ Profanity
Insult	27.14
Non-Insult	4.83
Directed Insult	31.12
Non-Directed Insult	5.79

Table 6: Breakdown of insult types for comments that contain profanity (by insult type).

containing an insult, we make the assumption that the profane term is used in the context of an insult.

First, we analyze the differences between comments that contain profanity and those that do not. Table 5 summarizes this breakdown. From Table 5, we see that, of all profane comments 58.66% contain an insult. Of all non-profane comments 16.19% contain an insult. These values differ significantly $\chi^2(1, N = 5265) = 652.464, p < .001$. That is, if a comment contains profanity, it is significantly more likely to also contain an insult. Similarly, 39.49% of all profane comments contain a directed insult while 8.15% of all non-profane comments contain a directed insult. This finding is also significant $\chi^2(1, N = 5017) = 561.473, p < .001$.

We also found significance for the inverse questions. That is, if a comment contains an insult, it is significantly more likely to contain profanity $\chi^2(1, N = 5265) = 1149.80, p < .001$ (see Table 6). Directed insults are also significantly more likely to contain profanity $\chi^2(1, N = 5017) = 639.143, p < .001$ (see Table 6). While these correlations do indicate that insults (and directed insults) and profanity are closely tied, it is still interesting to note that nearly 42% of all profane comments do not contain an insult at all. This indicates that there are uses of profanity within the corpus in a non-insulting context.

The next logical question is – in what context do these profane words occur if not in an insult? A manual investigation of this set of comments showed that nearly all occurred in negative ‘rants’ on the topic at hand. For example, the comments in Table 7 were labeled as profane comments that do not contain insults. Future work includes a more detailed analysis of comments that contain profane terms, yet no insult.

Next, we analyze differences in the context of profanity use between domains. Our method involves the profanity/insult co-occurrence measures used above to characterize the data set as a whole. In our analysis, comments in the domain of

“I’m done. I don’t give a F*** anymore. This Country is as good as gone. The Chosen ones and the Zionists won. Check out ‘Rules for Radicals’ and ‘The Protocols of the elders of Zion’ to see exactly what’s going on. Was nice while it lasted USA. Rest In Peace.....”
“So where are these f!@# jobs!?? You mean the 7.25 an hour job offered my daughter who has been managing a DQ for 3 years now? Or the temp clerical position that MAY go perm if the employer can make some money that they offered my wife.... by the way for \$10 an hour. How about the President getting paid \$40k a year and pay the bills on the white house and feed his family with that. Any excuse to raise friggen gas so some CEO can make a big salary is bull s***”
“Hey, Happy St. Patricks! Time to suck in new generations to drinking. Show them how fun and cultural getting sh*t faced on St Patricks Day is. Let them see the drunk tanks, impound lots, women shelters, ER’s, and morgues.”
“Ugh, not this b***** again.”

Table 7: Examples of comments labeled as profane, yet not containing an insult.

politics	% dir.		¬ politics	
	<i>insult</i>	<i>insult</i>	<i>insult</i>	<i>insult</i>
<i>Prof</i>	41.07	58.93	38.92	61.08
<i>¬Prof</i>	11.73	88.27	7.12	92.88

politics	% insult		¬ politics	
	<i>insult</i>	<i>¬insult</i>	<i>insult</i>	<i>¬insult</i>
<i>Prof</i>	62.99	37.01	57.14	42.86
<i>¬Prof</i>	22.84	77.16	14.25	85.75

Table 8: This table shows a comparison of the distributions of insults and directed insults among profane comments and among non-profane comments. We compare how these distributions differ between politics and non-politics comments.

politics were found to differ significantly from comments outside of the domain of politics in the distribution of insults and directed insults among profane and non-profane comments. Table 8 shows the distributions of insults and directed insults among profane and among non-profane comments. For insults, the breakdown differs significantly between politics and non-politics comments $\chi^2(3, N = 5265) = 66.75, p < .001$. For directed insults, it also differs significantly between politics and non-politics comments $\chi^2(3, N = 5017) = 33.038, p < .001$. Profanity use in the politics domain is tied more to insults and directed insults than in comments in other domains. That is, if a political comment contains profanity, it is more likely to

% of comments with	0 'rating down's	1+ 'rating down's
with profanity	36.64%	63.36%
without profanity	45.53%	54.47%
% of comments with	0 'rating up's	1+ 'rating up's
with profanity	22.02%	77.98%
without profanity	25.59%	74.41%

Table 9: A comparison of 'rating up's and 'rating down's in comments with and without profanity.

contain an insult or directed insult than a non-political comment.

HOW IS PROFANITY RECEIVED?

One might assume that profanity, like flames or personal insults, would discourage active user participation and engagement. To understand more about how profanity is received/tolerated, we looked to measures of the popularity of a comment within our data set. Most social news sites allow users to vote on comments in addition to stories, using features such as 'digg,' 'like,' 'thumbs up,' 'buzz up,' 'thumbs down,' and 'buzz down.' These features give us some additional popularity information about each comment. The social news site we studied allows users to both 'rate up' and 'rate down' each comment, and the number of 'rate up's and 'rate down's per comment are represented in our data set. We made the assumption that 'rate up's and 'rate down's could be interpreted as a measure of popularity or how much attention each comment received.

We divided our data set into comments labeled by MTurk workers as containing profanity, and those labeled as not containing profanity, and then looked at the difference in number of 'rating up's per profane comment and 'rating up's per non-profane comment (and similarly for 'rating down's). Table 9 shows the percent of profane comments with 0 and 1 or more 'rating down's, (and similarly for 'rating up's). For example, the upper left-most data point can be read as *36.64% of all profane comments received 0 rating down's*. We found that profane comments were significantly more likely to receive 'rate up' votes $\chi^2(1, N = 6354) = 3.990, p < .05$ and 'rate down' votes $\chi^2(1, N = 6354) = 18.965, p < .001$. Thus, profane comments are more popular or more widely read than non-profane comments. This confirms our intuition that passion (as interpreted by the use of profanity) towards a topic typically engenders either passionate agreement (compelling a user to 'rate up') or strong disagreement (causing a user to 'rate down').

LIMITATIONS

It is important to note several key limitations to our findings. First, the labeled data set on which we performed

this analysis was 6500 comments across ten domains. While there was a good variety of content in this data set, some domains had far fewer comments than others. As such, analysis beyond that accomplished in this paper will be done on a data set where the number of comments in each domain is balanced. Secondly, we have examined the nature of profanity use on just one user-generated content site. It would be inappropriate to generalize our findings beyond that site, as specific sites often attract distinct types of users who set up different norms about appropriate behavior. Little is known about how those norms are established and how they evolved. However, this study is a first step in establishing such an understanding.

CONCLUSIONS

In this paper, we made three primary contributions. The first concerned the state of current list-based profanity detection systems. Through an evaluation of the current state of the art in profanity detection, we argued that current systems do not suffice. The best performance we found from a list-based system was an f-measure of 0.457 (0.528 precision at 0.402 recall). This performance is quite poor for what is often underestimated as a simple task. Through the use of a data set of user-generated comments from a social news site, labeled by Amazon Mechanical Turk workers, we analyzed the salient differences between comments labeled as profane and not profane. This analysis exposed and emphasized our argument that current systems do not suffice because they fail to adapt to evolving profane language, misspellings (intentional or not), and profane terms disguised or partially censored by their author. The latter proved to be very prevalent in our finding of the most common features that distinguish profane from non-profane comments in our MTurk labeled data set (see Table 2).

Our second contribution is with regard to a major oversight of profanity detection systems – a lack of tailoring for specific communities. To establish the importance of this oversight, we provide evidence that communities not only use profanity with different frequencies, but also in different ways or contexts. In Table 4, we showed that comments in the politics community of Yahoo! Buzz were significantly more likely to contain profanity, insults, and directed insults (insults directed at other members of the community), than other communities. Similarly, we found that comments in the lifestyle community of Yahoo! Buzz were significantly less likely to include insults and directed insults, comments in the sports community of Yahoo! Buzz were significantly less likely to include directed insults, and comments in the business community of Yahoo! Buzz were significantly less likely to include insults than other communities. From this evidence, we conclude that different communities incite and permit differing amounts of profane language as these comments remained on the site and were not removed by a community manager or social moderation.

Next, addressing the context in which profanity is used, we find that overall, comments with profanity are significantly more likely to include an insult and a directed insult (see Table 5). While this is an intuitive conclusion, it also provided us with a method by which to analyze the differences between the contexts of profanity use in different domains. We analyzed how the propensity for a profane comment to include an insult differs by domain. Table 8 shows that profane comments in the politics domain are significantly more likely to contain insults and directed insults than in other domains. Combined with evidence that profanity is used at different frequencies in other domains, this drew us to conclude that profanity is used differently between communities.

Finally, we provided an analysis of how profanity is received. Using the standard community feedback mechanism of ‘rate up’ and ‘rate down’ we judged the popularity of comments with and without profanity. Surprisingly, we found that overall comments with profanity were both significantly more likely to receive ‘rate up’s and to receive ‘rate down’s.

FUTURE WORK

Following the conclusions drawn in this article, there are a few clear next steps with regard to moving beyond list-based profanity detection systems, and tailoring systems for specific communities.

First, since list-based profanity detection systems don’t suffice, future work involves building profanity detection systems from a machine learning point of view that take into account the context in which profane language is used. Learning the context, in addition to the actual profane words, has a greater potential for robustness, enabling the system to stand up to misspellings, disguised or partially censored words and evolving profane language. Similarly relevance feedback can be used to adapt and improve the model over time.

Secondly, since we established that profanity use and tolerance is very specific to a community, it is very clear that these systems will have to be developed or trained by each community. Future work involves building toolkits that allow this sort of tailoring. The use of Amazon Mechanical Turk and other low cost crowdsourcing mechanisms will prove crucial in labeling profanity in data sets from each community in order to train these machine learning systems.

Finally, we believe our results are most valuable as part of a larger investigation into the social nature of profanity and negative content within specific domains and user communities. In future studies we intend to extend our explorations of the social meanings of profanity and its context-specific use through qualitative interviews and survey studies. Furthermore, we expect that cross-site studies may be particularly revealing about the uses of profanity and possible context-specific approaches for

detecting it. In future work we hope to compare and contrast multiple data sets that share a topical domain (e.g. politics) but are drawn from several different sites.

REFERENCES

1. Boyd, D. and Marwick, A. Why Cyberbullying Rhetoric Misses the Mark. *The New York Times*, 2011. <http://www.nytimes.com/2011/09/23/opinion/why-cyberbullying-rhetoric-misses-the-mark.html>.
2. Callison-Burch, C. Fast, cheap, and creative: evaluating translation quality using Amazon’s Mechanical Turk. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, (2009).
3. Dinakar, K., Reichart, R., and Lieberman, H. Modeling the Detection of Textual Cyberbullying. *Proceedings of International AAAI Conference on Weblogs and Social Media, Workshop “Social Mobile Web,”* (2011).
4. Fleck, M.M., Forsyth, D.A., and Bregler, C. Finding Naked People. *European Conference on Computer Vision II*, (1996), 592 - 602.
5. Hammami, M., Chahir, Y., and Chen, L. WebGuard: a Web filtering engine combining textual, structural, and visual content-based analysis. *IEEE Transactions on Knowledge and Data Engineering* 18, (2006), 272-284.
6. Haninger, K. and Thompson, K.M. Content and Ratings of Teen-Rated Video Games. *JAMA: The Journal of the American Medical Association* 291, 7 (2004), 856 -865.
7. Ivory, J.D., Williams, D., Martins, N., and Consalvo, M. Good clean fun? A content analysis of profanity in video games and its prevalence across game systems and ratings. *Cyberpsychology & Behavior: The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society* 12, 4 (2009), 457-460.
8. Jacob, V., Krishnan, R., Ryu, Y., Chandrasekaran, R., and Hong, S. Filtering objectionable internet content. *Proceeding ICIS ‘99 Proceedings of the 20th international conference on Information Systems*, (1999).
9. Jacob, V.S., Krishnan, R., and Ryu, Y.U. Internet content filtering using isotonic separation on content category ratings. *ACM Transactions on Internet Technology* 7, (2007), 1-es.
10. Joachims, T. Text categorization with Support Vector Machines: Learning with many relevant features. In C. Nédellec and C. Rouveirol, eds., *Machine Learning: ECML-98*. Springer-Verlag, Berlin/Heidelberg, 1998, 137-142.
11. Jones, M.J. and Rehg, J.M. Statistical color models with application to skin detection. *INTERNATIONAL JOURNAL OF COMPUTER VISION* 46, (1999), 274--280.
12. Laboreiro, G., Sarmiento, L., Teixeira, J., and Oliveira, E. Tokenizing micro-blogging messages using a text classification approach. *Proceedings of the fourth*

- workshop on Analytics for noisy unstructured text data*, ACM (2010), 81–88.
13. Levenshtein, V.I. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10, 8 (1966), 707-710.
 14. Li, Q. Cyberbullying in Schools. *School Psychology International* 27, 2 (2006), 157 -170.
 15. Lou, J.K., Chen, K.T., and Lei, C.L. A collusion-resistant automation scheme for social moderation systems. *IEEE Consumer Communications and Networking Conference, 2009.*, (2009), 571 -- 575.
 16. Nguyen, D. and Rose, Carolyn. Language use as a reflection of socialization in online communities. *Proceedings of the Workshop on Language in Social Media.*, (2011).
 17. Sheng, V.S., Provost, F., and Ipeirotis, P.G. Get another label? improving data quality and data mining using multiple, noisy labelers. *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '08*, (2008), 614.
 18. Siersdorfer, S., Chelaru, S., Nejd, W., and San Pedro, J. How useful are your comments?: analyzing and predicting youtube comments and comment ratings. *Proceedings of the 19th international conference on World wide web*, (2010), 891–900.
 19. Sood, S.O., Churchill, E., and Antin, J. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology*, (2011).
 20. Spertus, E. Smokey: Automatic Recognition of Hostile Messages. *IN PROC. IAAI*, (1997), 1058--1065.
 21. Sukumaran, A., Vezich, S., McHugh, M., and Nass, C. Normative influences on thoughtful online participation. *Proceedings of the 2011 annual conference on Human factors in computing systems*, ACM (2011), 3401–3410.
 22. Tetreault, J.R., Filatova, E., and Chodorow, M. Rethinking Grammatical Error Annotation and Evaluation with the Amazon Mechanical Turk. *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, (2010), 45 - 48.
 23. Wang, J.Z., Li, J., Wiederhold, G., and Firschein, O. System for Screening Objectionable Images. *COMPUTER COMMUNICATIONS JOURNAL* 21, (1998), 1355--1360.
 24. Yin, D., Xue, Z., Hong, L., Davison, B., Kontostathis, A., and Edwards, L. Detection of Harassment on Web 2.0. *Proceedings of the Content Analysis in the WEB 2.0 (CAW2.0) Workshop at WWW2009*, (2009).
 25. Yoon, T., Park, S.-Y., and Cho, H.-G. A Smart Filtering System for Newly Coined Profanities by Using Approximate String Alignment. *Proceedings of the 2010 10th IEEE International Conference on Computer and Information Technology*, IEEE Computer Society (2010), 643–650.