International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2577-2589 www.ijaem.net ISSN: 2395-5252

Profanity Detector and Filter (Video)

Rahul Rawat
UG Student, IT, Maharaja Agrasen Institute of Technology, Delhi, India

---
Submitted: 10-06-2021 Revised: 20-06-2021 Accepted: 23-06-2021

ABSTRACT
As user-generated Web content increases, the amount ofinappropriate and/orobjectionablecontentalsogrows.Several scholarly communities are addressing how to detectandmanagesuchcontent:researchincomputervisionfocusesondetectionofinappropriateimages, naturallanguage processing technology has advanced to recognizeinsults.However,profanitydetectionsystemsremainflawed. Current list-based profanity detection systems havetwolimitations.First,theyareeasyto circumvent andeasilybecomestale–thatis,theycannotadapttomisspellings, abbreviations, and the fast pace of profaneslangevolution.Secondly,theyofferaone-sizefitsallsolution;theytypicallydonotaccommodatedomain,communityandcontextspecificneeds.However,socialsettingshavetheirownnormativebehaviors–whatisdeemedacceptableinonecommunitymaynot be inanother. In this paper, through analysis of comments from asocial news site, we provide evidence that current systemsare performing poorly and evaluate the cases on which theyfail.Wethenaddresscommunitydifferencesregardingcreation/tolerance of profanity and suggest a shift to morecontextually nuancedprofanity detectionsystems.

AuthorKeywords
Onlinecommunities,commentthreads,user-generatedcontent,negativity,community management,profanity.
ACMClassificationKeywords
H.5.3.Informationinterfacesandpresentation:GroupandOrganizationInterfaces.
GeneralTerms
Design,Experimentation,HumanFactors.

I. INTRODUCTION

Onlinecommunitiesareoftenplaguedwithnegative content – user-generated content that is negative in tone,hurtful in intent, mean, profane, and/or insulting. Negativecontent canbeproblematicforsiteswantingtoexpandtheir

user base, engage existing users, and foster a positive andcollaborativecommunity. Social contracts andnormativebehaviors, however, are unique to specific socio-technicalsystems.Whatisconsideredinappropriate ina givencontext is both site and community specific. On many sites,community managers are primarily responsible for the taskof removing inappropriate content. However, the flood ofuser-generated content on many sites quickly overwhelmscommunity managers'ability to effectively manageit.
Thedetectionofnegativecontentofmaliciousintent(personalattacksandinsults)inforumsandcommentstreams is a challenging and nuanced problem [20]. Recentwork in machine learning and natural language processinghas approached this task with varying degrees of success:withmaximalf-measuresof0.298fordetectionofharassment on Slashdot and 0.313 on MySpace from onestudy [24] and a maximal f-measure of 0.5038 for detectionofpersonalinsultsonanother[19].[1]Giventherecentattentiontothecomplexandsometimesgraveconsequences of cyberbullying [1], the ability to recognizeandpotentiallymitigateprofanityandotherformsofharmfulnegativityinuser-generatedcontentismoreimportant thanever[3,14].
Compared to the challenges of detecting malicious contentor spam, detection and removal of profanity is often thoughtto be an easier task. Most current approaches to profanitydetection

check new content against large lists of profaneterms. However, these systems are flawed in at least twomajorways.First,staticterm-listsquicklylosecurrencyandarerelativelyeasytocircumvent.Usersoftendisguiseorpartiallycensorprofanitybyreplacingone or moreletterswithpunctuationmarks(e.g.,“@ss,”“%@#$”).Misspellings,bothintentionalorunintentional(e.g.“biatch”,or“shiiiit”)andtheuseofslangwords(e.g.“assbite”)thatevolvequicklyandoftenhavelocalvariations also challenge profanity lists to be more thoroughandadaptivethantheycanreasonablybe.Thus,thesesystems face issues of recall; they are unable to catch mostcasesofprofanity.Secondly,list-basedapproachestoprofanity detection are a one-size-fits-all solution that doesnottakeintoaccountdifferencesincommunitynormsand

[1]F-measureisameasureofaccuracy,specificallytheharmonic meanof precisionandrecall.

practices.Afterall,whatconstitutesprofanitydiffers greatly based on the specific community and topic at hand.For example, in a forum about dog breeding, “bitch” is aterm of art that refers to a female dog, while in many othercontexts it is a profane term. Furthermore, sites for childrenhave a drastically different tolerance of profanity than thosefor adults.

Inthispaperwemakethreeprimarycontributionstoresearchonprofanitydetection.First,weaddressthestateof current list-based profanity detection systems. Do thesesystemssuffice?Inwhatcasesdotheyfail? Secondly,undertheassumptionthatamajoroversightin thesesystems is a lack of tailoring for specific communities, weexamine how profanity use differs between communities. Isprofanityusedmoreorlessinsomecommunities? Docertain communities use profanity in different ways? Andfinally, we explore the social context of profanity use indifferenttopicalcommunities.Howmightspecific communities receive profanity differently? In the sectionsthat follow, we examine these questions through analysis ofa data set of comments froma social news site.

II. BACKGROUND

As more and more of the web has grown to include user-generatedcontent,thedetectionandmanagementofinappropriateorobjectionablecontenthasbecomeanimportanttaskforwebsites.Onecommontechniqueissocial moderation, in which users themselves undertake thetask of identifying and flagging of profane or inappropriateresponses.Howeverthesesystemshavebeenonlymoderately successful, and suffer from potential collusion -flagging can be used to indicate disagreement or dislike of apostthatisnototherwiseinappropriateorprofane[15].Instead of relying on social moderation, recent proposalshave been made to automate the detection of inappropriateor abusivecontent.

Researchincomputervisionhasgivenmuchattentiontotherelatedissueofdetectinginappropriatevideosandimages.Advancesinthisspacehavelargelyincludedsystems that detect “too much skin” in images and videos[4,11,23].Othersystemsutilizetextualmetadata [8,9],while some combine the two; one such system, WebGuardhas reached 97.4% accuracy in detecting pornographic websites [5].

Whilemanywouldarguethattextualanalysisismore tractable than visual content analysis, this may be in partbecause of a general misunderstanding about how difficultthe problem of profanity detection is in real-world contexts.Furthermore,texthasavisualelementthatissociallyunderstood.Expressiveformssuchasemoticonsand“ASCII art” use visual properties of text, punctuation marksandsymbolstomimiclexicalunitsandthusconveymeaning, denote profanity and circumvent automatic filters.Suchvisual-for-textualsubstitutionisbestillustratedthroughexamples suchas theuse of “@” in “@ss”..

Becauseofthesemisunderstandings,perhaps,comparativelylittleresearchhasfocusedondetectinginappropriatetextinuser-generatedcontentsystems.Asmentioned above, two groups have built systems to detectinsults and harassment in online forums [19,24] and anotherhas focused on cyberbullying of teens [3], but even fewerhave addressed the identification of profanity. Yoon et al.built a system to detect newly coined profanities in Korean,inanattempttoimproveuponthefailureoflist-basedsystems to evolve along with profane language [25]. Due tothe target audience being children, some have analyzed thecontentofvideogamewebsitesandvideogamesthemselves to verify that presented content meets ratingsstandards[6,7].However,workinthisareadoesnotgenerallystrive for automatedanalysis.

Advancing our ability to detect and remove profanity couldhave several significant, positive social consequences. Thegrowthofcollaborativeinformationproductssu

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2577-2589 www.ijaem.net ISSN: 2395-5252

chasWikipedia, Yahoo! Answers, and Stack Overflow rely ontheprovisionofinteractionenvironmentsthatares upportive, productive, and meet the specific needs of theiruser communities. Open-source software projects also relyonemaillistsandforumstosupportthenecessary communitybuilding,coordination,anddecision-makingprocesses.

No automated system, by itself, can appropriately filter andmanage ongoing discourse and interaction so that it meetsthe needs of a particular topic, domain, or user community.Indeed,researchhasillustratedtheimpo rtantroleofestablishedcommunitymembersforimp licitlyandexplicitly communicating language norms to new members[16].Theenforcementofthesenormsisofte nadhoc,however.Inlargesystems,thesheervolume ofcontentmeansadhocstrategiesoftenleavealargea mountofprofane or inappropriate user-generated content undetected.The existence of such content can actually fight against thepositive influence of community managers and long-timeparticipantsbysettingabadprecedentthatcom municatesto new users that profanity and other negative content isacceptable [21]. Automated systems that help communitymanagers,moderators,andadministrat orstomanagetheflood of user-generated content in these environments couldhelp to promote more productive large-scale collaborationandthus more valuable informationproducts.

III. DATASET

Socialnewssites(e.g.,reddit.com,digg.com)typical lyallowuserstopostlinkstostoriesofinterest,voteon contributed stories, and most important to the present study,comment on stories and respond to others' comments. Ourdata set is the complete set of user-contributed commentsover a three-month period (March to May 2010) to Yahoo!Buzz,asocialnewscommentingsitethatisno longeractive.Ourdatasetcontains1,655,131comm entsdistributedamong168,973distinctthreads.Inad ditiontothecommentitself,ourdatasetcontainsmeta informationabout each comment including the time posted and whichnews story the comment is in response to; this informationcan be combined to reproduce the comment thread. We alsohaveinformationabouteachnewsstoryincludin gitscountryoforigin,language,andcategory(i.e.,po litics,entertainment).Moreinformationaboutthisda taset,including the distributions of comment lengths, commentsperuserandcommentsper thread can be found in[19].

Coding

In order to generate a data set describing the presence ofprofanity,insults,andtheobjectsoftheinsults,wee mployed Amazon Mechanical Turk (MTurk). MTurk is anonlinelabormarketinwhichrequesterspostjobsth atcanbe easily decomposed into a large number of small tasks.MTurkworkers("Turkers")arepresentedwit hashortdescription of available tasks, and then choose which taskstocomplete.Individualtaskstypicallytakebet ween5and 20secondstocompleteandworkersaregenerally paidabout 5centsfor eachtask.

Recent studies have suggested that using MTurk for similarcontentanalysistaskscanbebothfasterandm oreeconomicalthanusingdedicatedtrainedraters[2 2].Furthermore, several studies have illustrated that combiningthe work of multiple Turkers on the same task can producehighqualitycontentanalysisresults,evenw hen somecodersdo notagree(i.e.thecoded datais"noisy")[2,17].

We selected a random sample of 6500 comments spanningall categories. Comments that were likely to be too short tomeaningfully interpret or too long to quickly process werenotsampled:werestrictedsamplingtothe2ndan d3rdquartiles for overall comment length (between 73 and 324characters long).

Each worker was shown one comment at a time. For eachcomment,theywereaskedtoanswerthefollowi ngquestions:

"Doesthismessagecontainanycontentyouwouldde scribeas'profanity?'(includingprofanitythatisdisg uisedormodifiedsuchas@ss,s***,andbiatch)(Yes/ No)

Thinking about the intent of the comment's author, does themessagecontainwhatyouwoulddescribeasa direct"insult?" (Yes/No)

Inyouropinion,istheinsultdirectedattheauthorofap reviouscomment?(Yes/No/Unsure)

Finally, beyond the requirement of a consensus thresholdfrommultiplecoders,wealsoemployeda 'gold data'model to improve label quality. Gold data were a set ofcomments for which the 'correct' labels (answers to theabove three

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2577-2589 www.ijaem.net ISSN: 2395-5252

questions) were designated by the researchersprior to the labeling task. If a Turker mislabeled one of thegold comments, he was shown a short explanation for thecorrect answer. In this way the gold data functioned as alightweighttrainingprogram.Inaddition,ifanyTurkerincorrectlylabeledtoomanyofthegoldcomments,their labels were removed from the data set and they were barredfromlabelinganyfurthercomments.Allthree authorsindependentlyjudgedthe'correct'labelsfor goldcomments. For most gold comments the authors agreed onan answer which was likely to be self-evident. For example,the following comment was judged to contain profanity butnoinsult:

"Nowthey'lljustreleasethemsotheycandothesamethingtomorrow. Mine the Effen Border. Use the Natl. Guard to patrolour borders with extreme prejudice. Fences don't work. They'lltunnel under them or use ladders. Show the Drug Cartels that wemeanbusinessandthissh!twillcease."

In other cases, however, the primary purpose of the goldcomment was to draw attention to the desired aspect of thecomment.Forexample,inthefollowingcommentitis,arguably,notpossibletoconclusivelydeterminewhoisbeinginsulted:

"Hot off the presses...straight fom their leader oweduma's mouth.Thestateofoureconomydeservesattention. Thisguymustliveinabunker.Wakeupyouliberalloser!Theeconomysucks."

Overthecourseofapproximately5days,221MTurk workersprovided25,965judgmentson6500comments.Following the model suggested by Sheng and colleagues[17],weemployedmultiplecodersforeachitem.Asaresult,eachitemwasratedbyaminimumofthreeraters.We adopted a simple consensus model on the labels. Toensurelabelingaccuracy,ourfinalprofanitylabeleddataset only includes those comments for which at least 66% oflabelers agreed on the profanity label. Similarly, our finalinsult and insult object labeled data sets only include thosecomments for which at least 66% of labelers agreed on theinsultorinsultobjectlabel.Thismethodresultedinadifferent N depending on the focal phenomena (profanity,insult,orinsultobject).Forexample,onehundredandforty-sixcomments(2.2%)weredroppedfromthefinalprofanitylabeleddatasetbecauseratersdidnotreachconsensus ontheprofanitylabel.

DOCURRENTPROFANITYDETECTIONSYSTEMSSUFFICE?

Thestandardapproachtoprofanitydetectioninonlinecommunitiesistocensororfiltertext basedonlistsofprofaneterms.Whenuser-generatedtextcontainslistedwords,thosewordsortheentirecontributionmay beflaggedforrevieworautomaticallyremoved.Someprofanitylistsaresharedbetweenmultiplesites,and administrators contribute additional terms as they becomeprevalent or problematic. In order to test the efficacy of thisapproach, we downloaded a shared profanity list from thesitephorum.organdbuiltasimplesystemthatflagsacomment as profane if it contains any of the words on thephorum.orglist.[2]

[2] At the time of our analysis (July 1, 2011), the phorum.orglist contained120profaneterms.

noswearing.com w/stemming	0.528	0.402	0.457	0.907
noswearing.com orphorum.org w/stemming	0.490	0.412	0.448	0.902
noswearing.com orphorum.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 detectionsystems.

system is included as a base line system; itrandomly labels

ass	asses	bullsh!t
s***	pussy	sob
f*****	cr	bitch
azz	stfu	schit
bastard	dumbass	sh
sh!t	$hit	r**h***
a******	f***	c***
nr	m*****f*****	phag!
b*******	f*cking	fing
sh*t	@ssp	goddamnbull$hi
a$$	ussies	t

Table2:Topwordsdistinguishingprofanefromnon-profanecomments inthedataset.

Asnotedabove,list-basedsystemsoftensuffer(byidentification/"recall"measures)asprofanelanguageevolves over time with slang and Internet abbreviations. Assuch, we downloaded a second list of profane terms fromnoswearing.com.Thissitehostsalistofcommunitycontributed profane terms. This list evolves over time withuser contributions and is larger than the phorum.org list.[3]While both lists contain traditional profane terms, they alsocontaininappropriatetermssuchasracialslursandvulgarities.

In another attempt to improve recall, we employ a stemmer.Beyond simply looking for the presence of a word on aprofanity list, the stemmer allows the system to see if anywords in a comment have a shared stem with any word on aprofanity list. To evaluate the efficacy of list-based methodswe built several systems that employed the two lists andstemming in various combinations. For each system, weaverageitsperformanceover5trialsof10-foldcrossvalidation on our 6500 profanity-labeled comments fromYahoo! Buzz. While the data set as a whole contains 6500comments,6354meetthe66%labelingconsensusacrossthe MTurk labelers for the profanity label. Of those 6354,595 (9.4% of the corpus) are positive cases, meaning thatthey contain profanity. All systems are evaluated based ontheirprecision(ameasureoffalsepositives),recall (ameasure of false negatives), f-measure (f1 – the harmonicmeanofprecisionandrecall)andaccuracy –thisisthestandard array of evaluation metrics for systems of this type[10,18,24].

TheperformancesofallsystemsaresummarizedinTable1,sorted indescending orderoff-measure.Therandom

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

commentsasprofaneornot.Similarly,the weightedrandom system labels comments randomly, weighted by thedistributionofprofane/non-profanecommentsinthetrainingset.Theperformancesofthesesystemsareincluded for comparison purposes, though they of courseapproach the theoretical random baselines. The remainingsystemsarelist-basedapproaches,basedonthelistsgathered from phorum.org and from noswearing.com.In anattempttoreachhigherrecall,wecreated additionalsystemsthatmarkedatermasprofaneifitappearedineither one of the two lists.Finally, in some systems wecombinedwordlists withstemming.

Whileapeakaccuracyof0.913seemspromising,recallthat 9.4% of the comments in our corpus contain profaneterms.Forthistestingdata,ifonebuiltasystemthat,givenacomment,alwaysreturnedanegativeclassification(indicating that the comment does not contain profanity), itwould have an accuracy of 0.906 as 90.6% of the testingcorpusiscommentsthatdonotcontainprofanity.Therefore,f1,precision,andrecallaremuchmore descriptiveevaluationmetrics.Asseenintable1,peakperformance of the list-based approaches is reached usingthe profane terms list from noswearing.com combined withastemmingsystem.Thissystemdetected40.2% oftheprofanity cases at 52.8% precision. Based on our results, wemust conclude that even the best of these list and stemmerbasedsystemswouldnotperformwellatdetectingandremoving profanityinuser-generatedcomments.

WHYDOLIST-BASEDAPPROACHESPERFORMSOPOORLY?

Aswehavealreadydiscussed,list-basedapproachesperformpoorlybecauseofthreeprimaryfactors:misspellings (both intentional and not), the context-specificnatureofprofanity,andquicklyshiftingsystemsofdiscoursethatmakeithardtomaintainthoroughandaccuratelists.Toexemplifytheseproblems,we analyzedthewordsthatmostcommonlydistinguish profanefrom

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2577-2589 www.ijaem.net ISSN: 2395-5252

non-profane comments in our MTurk profanity labeled dataset.Thetopwords,seeninTable2,weresortedin descendingorder byx, calculatedas follows:

(posFeatureCt/TotalPosFeatures)
x□ ———————————————
(negFeatureCt/TotalNegFeatures)

whereposFeatureCtisthenumberoftimesthewordoccurred in positive (profane) comments, negFeatureCtisthecorrespondingvaluefornegative (non-profane)

Figure1:Examplesoftwotweets,illustratingtheuseof#,@and http://bit.ly.

comments,TotalPosFeaturesisthesumofall featurecountsacrossallwordsinthepositivecomments,andTotalNegFeaturesisthecorrespondingvaluefor the

Context
Count ofOccurrencesof '@'
% of@usage
%of fulldata set inthiscontext

negativecomments.Thelattervaluesareincludedto adjust
for differencesin the profane and non-profane corpora
profane terms(i.e., 'ass,' 'bastard,' 'asses,' 'pussies,'
'pussy,' 'dumbass,' 'goddamn' and 'bitch'). There are sixinstances of slang abbreviations for profane terms – 'sob,''nr,' 'cr,' 'sh,' 'f'ing,' and 'stfu.' The remaining nineteentermsaredisguisedorauthorcensoredprofanity(e.g.,'bullsh!t,''azz,''f*****').Thus,alist-basedprofanitydetection system, such as the ones evaluated in the previoussection, would fail to catch twenty-five of the top thirty-three profane terms (76%) used in our data set. While thesewordscould, ofcourse,beaddedtoaprofanity list forfuture detection via a list-based system, there are countlessfurtherwaystodisguiseorcensorwords.Thismakesrepresenting themin alist asignificantchallenge.

Asfurtherevidenceofhowwidespreadtheparticular problem of disguised or partially censored profanity is, weanalyze use of one specific character, the @ symbol. ThepopularityofTwitterandothersocialmediahave resultedin adaptations and specializations of language for onlinecommunication[12].Justastextmessaginghasanestablishedshareddictionaryofacronyms,social mediashare some community established abbreviations that allowusers to pack more content into short messages. One suchabbreviationisthe'@'symbol.Whenauser writes'@rick', they are directing their message to 'rick', but in apublic medium. The '@' symbol provides a short and easymechanism for directing public comments towards specificindividuals,butalsohelpstobridgethegapbetweendirected and undirected interaction in computer-mediatedcommunication.Forexample:"@xeelizCheckthisout!http://yhoo.it/rq1y2u#NBAFinals."Twomoreexampletweets are shown in Figure 1. The top tweet from edchi,shows a use of the #. The bottom tweet from kevinmarksincludes a use of the @ symbol, indicating that this tweet isdirectlyaddressingthe userfeliciaday.

To study how the @ symbol is used within our completedatasetof1.65millioncomments,welookedatall

Context	Count ofOccurrencesof '@'	% of@usage	%of fulldata set inthiscontext
Emailaddress	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%

Table3:Analysisof'@'symbolusagewithinthedataset.

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2577-2589 www.ijaem.net ISSN: 2395-5252

comments that contained an instance of the symbol. Wefound 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.Somecommentscontainemailaddresses(e.g."john@somecompany.com") or direct readers to a website(e.g. "@ www.cnn.com"). We also found that commentsoften use the '@' symbol to disguise (e.g. "@ss") or censor(e.g. "@%#$") profanity – one of the very problems thatplaguestheprofanity detection systemsdescribed above.

To explore the multiple uses of the '@' symbol we built arulebasedsystemusingregularexpressions. Classifying'@'usageaswithinemailorwebaddressesiseasilyaccomplishedwithregularexpressions,however,automaticallydeterminingthat'@ss'isprofanitywhile'@john'isconversationalisamoredifficulttask.Weemployacorpusofprofaneterms(thelistsfromphorum.organdnoswearing.com),alongwitha tooltocalculate 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.Whenatermcontainsthe'@'symbol,in order todetermineifitisprofanity,wechecktoseeiftheLevenshtein edit distance between the term and any knownprofane term is equal to the number of punctuation markspresent in the term. For example '@ss' has one punctuationmark ('@') and has an edit distance of one from the profaneterm 'ass.' '$%#@' has four punctuation marks and has aneditdistanceoffourfromanyfourletterprofane term.Using this approach, we have a very high precision tool thattakes a term containing the '@' symbol and determines if itis a profane term (either disguised or censored). The recallofthistool isonlyasgoodasour listof profaneterms.

	Profanity		**Insult**		**DirectedInsult**	
Category	**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^2values are the results of thecomparisonofprofanity,insult,anddirectedinsultfrequencywithinagivencategorytothefrequencyacrossallothercategories. Throughoutthispaper,reportedsignificancevaluesareBonferroniadjustedwheretherearemultiplecomparisons.

***** p≤.001, **p≤.01, *p≤.05, † p≤.1**

Using this tool, we labeled all uses of the '@' symbol in ourcorpuswith'emailaddress,''webaddress,''profanity,''conversational,'or'other'(forinstancesthatwerenotprofanitybutalsodidnotappeartotaketheformofaconversational usage of '@'). The results can be seen inTable3.First,notethatonly0.67%(11,092)ofallcommentsinthedataset(1,655,131)containan'@'symbol. Within this set, 39.9% of '@' usage was within thecontext of a censored or disguised profane term, while only24.9%of'@'usagesappearin a conversationalcontext.

Nearly 40% of all occurrences of the @ symbol came in theformofdisguisedorauthor-censoredprofanity.The@symbol is just one of many punctuation marks that could beused to disguise profanity. Moreover, the @ symbol is onethat is thought to be commonly used in social media in aconversational manner, yet an astonishing 40% of its useswithin our data come in the form of disguised profanity.This is likely to be a conservative estimate, as it is knownthat list-based measures suffer in recall, as shown in theprevious section.

HOWFREQUENTLYISPROFANITYUSED?

In addition to facing issues of recall, list-based approachesare a one-size fits all approach that do not take into accounthow profanity is used within different domains, contexts,and communities. Through our MTurk labeled data set, weexplore the use of profanity in comments on news stories inordertounderstandmoreaboutthefrequencyandcontextof profanityuse andhowit is received.

First, we examine the prevalence of profanity in differenttopical domains. Dividing our 6500 labeled comments bydomain of the article they are in reference to, we see thatcommentsonpoliticalstoriescontainmoreprofanity,more insults, and more directed insults (directed at authors ofprevious comments) thaninanyother domain.

Table4showsthedistributionofprofanity,insultsanddirected insults in comments within the different domains.ToavoidthepossibilityofTypeIerrorweappliedtheconservativeBonferroniadjustmenttoallsignificancevalues reported. For clarity, the first value in Table 4 can beread as 10.7% of political comments contain profanity. Aspreviouslydiscussed,theNdiffersbetweenprofanity,insults, and directed insults because, for each we use onlyitems for which coders reached consensus. For profanity, Nis 4409, for insults N is 4177 and for directed insults N is3974. Each comment in the table 4 analyses was labeledwith one of the 10 categories shown. Profanity usage inpolitical comments is significantly more common than inother comments. Political comments contain significantlymore insults and directed insults than in other domains. Ontheotherendofthespectrum,thelifestyle commentscontain significantly fewer insults and directed insults thanother domains. The business domain also held significantlyfewerinsults,whilethesportsdomainhadsignificantlyfewer directed insults. As expected, from these data we canconclude that different domains of news story incite varyingamounts of profane language and use of insults (generalinsultsaswellasthosedirectedatothercommunitymembers).

INWHATCONTEXTISPROFANITYUSED?

Giventhat our dataset containsinsult andinsult objectlabels in addition to profanity labels, these labels will beutilized as a measure of context. To further understand howprofanity is used within our data set, we investigate the co-occurrence of the ‘comment contains profanity’ label with‘comment contains insult’ and ‘comment contains directedinsult’labels.Ifa commentislabeled asboth profane and

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

Table5:Ofallprofaneandnon-profanecomments,thistablepresentsthebreakdownofthosethatcontainaninsultoradirected insult.

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

Table6:Breakdownofinsulttypesforcommentsthatcontainprofanity (byinsulttype).

containinganinsult,wemaketheassumptionthatthe profanetermis usedinthe context ofaninsult.

First, we analyze the differences between comments thatcontain profanity and those that do

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2577-2589 www.ijaem.net ISSN: 2395-5252

not. Table 5 summarizesthis breakdown. From Table 5, we see that, of all profanecomments58.66%containaninsult.Ofallnon-profanecomments 16.19%contain an insult. These values differsignificantly □^{2}(1, N = 5265) = 652.464, p < .001. That is, ifa comment contains profanity, it is significantly more likelyto also contain an insult. Similarly, 39.49% of all profanecomments contain a directed insult while 8.15% of all non-profane comments contain a directed insult. This finding isalso significant□^{2}(1, N =5017)=561.473,p <.001.

We also found significance for the inverse questions. Thatis, if a comment contains an insult, it is significantly morelikelytocontainprofanity□^{2}(1,N=5265)=1149.80,p< .001 (see Table 6). Directed insults are also significantlymore likely to contain profanity □^{2}(1, N = 5017) = 639.143,p < .001 (see Table 6). While these correlations do indicatethat insults (and directed insults) and profanity are closelytied, it is still interesting to note that nearly 42% of allprofanecommentsdonotcontainaninsultatall.Thisindicatesthattherearеusesofprofanitywithinthecorpusinanon-insultingcontext.

The next logical question is– in what context do theseprofanewordsoccurifnotinaninsult?Amanual investigation of this set of comments showed that nearly alloccurredinnegative‘rants’onthetopicathand.Forexample, the comments in Table 7 were labeled as profanecommentsthatdonotcontaininsults.Future workincludesa more detailed analysis of comments that contain profaneterms, yetnoinsult.

Next, we analyze differences in the context of profanity usebetween domains. Our method involves the profanity/insultco-occurrence measures used above to characterize the datasetasa whole.In ouranalysis,commentsin the domain of

"I'm done. I don't give a F*** anymore. This Country isas good as gone. The Chosen ones and the Zionists won.Check out 'Rules for Radicals' and 'The Protocals oftheelders of Zion' to see exctly whats going on. Was nicewhileitlasted USA.RestIn Peace......."
"So where are these f!@# jobs!?? You mean the 7.25 anhourjob offered my daughter who has been managing aDQ for 3 years now? Or the temp clerical position thatMAY go perm if the employer can make some moneythatthey offered my wife bythe way for$10anhour. How about the President getting paid $40k a year and paythe bills on the white house and feed his family with that.Any excuse to raise friggen gas so some CEO can make abigsalaryisbull s***"
"Hey,HappySt. Patricks!Timeto suck in newgenerations to drinking. Show them how fun and culturalgetting sh*tfaced on St Patricks Day is. Let them see thedrunk tanks, impound lots, women shelters, ER's, andmorgues."
"Ugh,notthisb*******again."

Table7:Examplesofcommentslabeledasprofane,yetnotcontaining aninsult.

politics			**□politics**	
	% dir.insult	**%** □ dir.insult	% dir.insult	**%**□ dirinsult
Prof	41.07	58.93	38.92	61.08
□ Prof	11.73	88.27	7.12	92.88
politics			**□politics**	

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2577-2589 www.ijaem.net ISSN: 2395-5252

	% insult	% insult	% insult	% insult
Prof	62.99	37.01	57.14	42.86
Prof	22.84	77.16	14.25	85.75

Table8:Thistableshowsacomparisonofthedistributionsofinsultsanddirectedinsultsamongprofanecommentsandamongnon-profanecomments.Wecomparehowthesedistributionsdifferbetweenpoliticsandnon-politicscomments.

politics were found to differ significantly from commentsoutsideofthedomainofpoliticsinthedistribution ofinsults and directed insults among profane and non-profanecomments. Table 8 shows the distributions of insults anddirectedinsultsamongprofaneandamongnon-profanecomments. For insults, the breakdown differs significantlybetweenpoliticsandnon-politicscomments $\chi^2(3,N=5265)=66.75,p<.001$.Fordirectedinsults,italsodifferssignificantlybetweenpoliticsandnon-politicscomments

1+ 'ratingup's

natureofprofanityuseonjustoneuser-

withprofanity 22.02% 77.98%

$\chi^2(3, N = 5017) = 33.038, p < .001$. Profanity use in thepolitics domain is tied more to insults and directed insultsthan in comments in other domains. That is, if a politicalcommentcontains profanity, it is more likelyto

somedomainshadfarfewercommentsthanothers. Assuch,analysisbeyondthataccomplishedinthispaperwillbedoneonadatasetwherethenumberofcommentsin eachdomainisbalanced.Secondly,wehaveexaminedthe

generatedcontentsite.Itwouldbeinappropriatetogeneralizeourfindings

withoutprofanity 25.59% 74.41%

%ofcommentswith	0'rating down's	1+ 'ratingdown's
withprofanity	36.64%	63.36%
withoutprofanity	45.53%	54.47%
%ofcommentswith	0'rating up's	1+ 'ratingup's
withprofanity	22.02%	77.98%
withoutprofanity	25.59%	74.41%

Table9:Acomparisonof'ratingup'sand'ratingdown'sincommentswithandwithoutprofanity.

containaninsultordirectedinsultthananon-politicalcomment.

HOWISPROFANITYRECEIVED?

One might assume that profanity, like flames or personalinsults,woulddiscourageactiveuserparticipationandengagement. To understand more about how profanity isreceived/tolerated, we looked to measures of the popularityof a comment within our data set. Most social news sitesallowuserstovoteoncommentsinadditionto stories,using features such as 'digg,' 'like,' 'thumbs up,' 'buzz up,''thumbs down,' and 'buzz down.' These features give ussomeadditionalpopularityinformationabouteachcomment. The social news site we studied allows users toboth'rateup'and'ratedown'eachcomment,andthenumber of 'rate up's and 'rate down's per comment arerepresented in our data set. We made the assumption that'rateup'sand'ratedown'scouldbeinterpretedasameasure of popularity or how much attention each commentreceived.

We divided our data set into comments labeled by MTurkworkers as containing profanity, and those labeled as notcontaining profanity, and then looked at the difference innumberof'ratingup'sperprofanecommentand'ra

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2577-2589 www.ijaem.net ISSN: 2395-5252

tingup's per non-profane comment (and similarly for 'ratingdown's). Table 9 shows the percent of profane commentswith 0 and 1 or more 'rating down's, (and similarly for'rating up's). For example, the upper left-most data pointcan be read as 36.64% of all profane comments received 0ratingdown's.Wefoundthatprofanecommentswe resignificantly more likely to receive 'rate up' votes χ^2(1, N =6354)=3.990,p<.05and'ratedown'votes χ^2(1,N= 6354)=18.965,p<.001.Thus,profane comments aremorepopularormorewidelyreadthannon-profanecomments.Thisconfirmsourintuitionthatp assion(asinterpreted by the use of profanity) towards a topic typicallyengenderseitherpassionateagreement(co mpellingauserto 'rate up') or strong disagreement (causing a user to 'ratedown').

IV. LIMITATIONS

Itisimportanttonoteseveralkeylimitationsto ourfindings.First,thelabeleddatasetonwhichweper formed
beyond that site, as specific sites often attract distinct
typesofuserswhosetupdifferentnormsaboutapprop riatebehavior.Littleisknownabouthowthosenorms areestablished and how they evolved. However, this study is afirststepinestablishingsuch anunderstanding.

V. CONCLUSIONS

Inthispaper, wemadethree primary contributions.
Thefirstconcernedthestateofcurrentlist-basedprofanitydetectionsystems.Throughanevalu ationofthe currentstate of the art in profanity detection, we argued that currentsystemsdonotsuffice.Thebestperformance we foundfrom a list-based system was an f-measure of 0.457 (0.528precisionat 0.402recall). This performance is quite poorfor what is often underestimated as a simple task. Throughthe use of a data set of user-generated comments from asocialnewssite,labeledbyAmazonMechanicalTu rkworkers,weanalyzedthesalientdifferencesbetwe encomments labeled as profane and not profane. This analysisexposed and emphasized our argument that current systemsdo not suffice because they fail to adapt to evolving profanelanguage,misspellings(intentionalornot),a ndprofaneterms disguised or partially censored by their author. Thelatter proved to be very prevalent in our finding of the mostcommon features that distinguish profane from non-profanecommentsinourMTurklabeled dataset(seeTable 2).
Our second contribution is with regard to a major oversightofprofanitydetectionsystems–alackoftailoringforspecific communities. To establish the importance of thisoversight, we provide evidence that communities not onlyuseprofanitywithdifferentfrequencies,butalso indifferentwaysorcontexts.InTable4,weshowedth atcomments in the politics community of Yahoo! Buzz weresignificantly more likely to contain profanity, insults, anddirected insults (insults directed at other members of thecommunity), than other communities. Similarly, we foundthat comments in the lifestyle community of Yahoo! Buzzwere significantly less likely to include insults and directedinsults, comments in the sports community of Yahoo! Buzzwere significantly less likely to include directed insults, andcomments in the business community of Yahoo! Buzz weresignificantlylesslikelytoincludeinsultsthanot hercommunities.Fromthisevidence,weconcludeth atdifferentcommunitiesinciteandpermitdifferinga mountsof profane language as these comments remained on the siteand were not removed by a community manager or socialmoderation.
Next, addressing the context in which profanity is used, wefind that overall, comments with profanity are significantlymore likely to include an insult and a directed insult (seeTable5).Whilethisisanintuitiveconclusion,ital soprovideduswithamethodbywhichtoanalyzethed ifferencesbetweenthecontextsofprofanityuse indifferent domains. We analyzed how the propensity for aprofane comment to include an insult differs by domain.Table8showsthatprofanecommentsin the politicsdomain are significantly more likely to contain insults anddirectedinsultsthaninotherdomains.Combined withevidence that profanity is used at different frequencies inother domains, this drew us to conclude that profanity isuseddifferentlybetweencommunities.
Finally,weprovidedananalysisofhowprofanityisre ceived.Usingthestandardcommunityfeedbackmec hanismof'rateup'and'ratedown'wejudgedthepop ularityofcommentswithandwithoutprofanity.Surp risingly,wefoundthatoverallcommentswithprofan itywerebothsignificantlymorelikelyto receive'rateup's andtoreceive'ratedown's.

VI. FUTUREWORK

Following the conclusions drawn in this article, there are afew clear next steps with regard to moving beyond list-based profanity detection

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2577-2589 www.ijaem.net ISSN: 2395-5252

systems, and tailoring systems forspecificcommunities.

First,sincelist-basedprofanitydetectionsystemsdon'tsuffice, future work involves building profanity detectionsystems from a machine learning point of view that takeinto account the context in which profane language is used.Learningthecontext,inadditiontotheactualpr ofanewords, has a greater potential for robustness, enabling thesystem to stand up to misspellings, disguised or partiallycensored words and evolving profane language. Similarlyrelevance feedback can be used to adapt and improve themodel overtime.

Secondly,sinceweestablishedthatprofanityuseand toleranceisveryspecifictoacommunity,itisverycle arthat these systems will have to be developed or trained byeachcommunity.Futureworkinvolvesbuilding toolkitsthatallowthissortoftailoring.TheuseofAm azonMechanicalTurkandotherlowcostcrowdsourc ingmechanisms will prove crucial in labeling profanity in datasets from each community in order to train these machinelearningsystems.

Finally, we believe our results are most valuable as part of alarger investigation into the social nature of profanity andnegativecontentwithinspecificdomainsanduse rcommunities.Infuturestudiesweintendtoextendo urexplorationsofthesocialmeaningsofprofanityan ditscontext-specificusethroughqualitativeinterviewsandsurve ystudies.Furthermore,weexpectthatcross-sitestudiesmaybeparticularlyrevealingabouttheus esofprofanityandpossiblecontext-specificapproachesfor detectingit.Infutureworkwehopetocompare andcontrast multiple data sets that share a topical domain (e.g.politics)but are drawnfromseveraldifferent sites.

REFERENCES

[1]. Boyd,D. andMarwick, A. WhyCyberbullyingRhetoric 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 translationquality usingAmazon'sMechanical Turk. Proceedings of the 2009 Conferenceon Empirical Methods inNatural LanguageProcessing, (2009).

[3]. Dinakar, K., Reichart, R., and Lieberman, H. ModelingtheDetectionofTextualCyberbullying.Proceedingsof International AAAI Conference on Weblogs andSocialMedia,Workshop"SocialMobile Web,"(2011).

[4]. Fleck, M.M., Forsyth, D.A., and Bregler, C. FindingNaked People. European Conference on ComputerVision II, (1996),592- 602.

[5]. Hammami, M., Chahir, Y., and Chen, L. WebGuard: aWeb filtering engine combining textual, structural, andvisual content-based analysis. IEEE Transactions onKnowledgeandDataEngineering18,(2006),272-284.

[6]. Haninger, K. and Thompson, K.M. Content andRatings of Teen-Rated Video Games. JAMA: TheJournal of the American Medical Association 291, 7(2004), 856-865.

[7]. Ivory,J.D.,Williams,D.,Martins,N.,andConsalvo,

[8]. M. Good clean fun? A content analysis of profanity invideo games and its prevalence across game systemsand ratings. Cyberpsychology & Behavior: The Impactof the Internet, Multimedia and Virtual Reality onBehaviorand Society 12, 4(2009), 457-460.

[9]. Jacob, V., Krishnan, R., Ryu, Y., Chandrasekaran, R.,and Hong, S. Filtering objectionable internet content.Proceeding ICIS'99Proceedings ofthe20thinternational conference on Information Systems,(1999).

[10]. Jacob, V.S., Krishnan, R., and Ryu, Y.U. Internetcontent filtering using isotonic separation on contentcategory ratings. ACM Transactions on InternetTechnology 7, (2007),1-es.

[11]. Joachims, T. Text categorization with Support VectorMachines: Learning with many relevant features. In C.Nédellec and C. Rouveirol, eds., Machine Learning:ECML-98. Springer-Verlag, Berlin/Heidelberg, 1998,137-142.

[12]. Jones, M.J. and Rehg, J.M. Statistical color modelswith application to skin detection. INTERNATIONALJOURNAL OF COMPUTER VISION 46, (1999), 274--280.

[13]. Laboreiro,G.,Sarmento,L.,Teixeira,J.,and Oliveira, E. Tokenizing micro-blogging messages using a textclassificationapproach.

Proceedingsofthefourth workshop on Analytics for noisy unstructured textdata, ACM(2010),81–88.
[14]. Levenshtein, V.I. Binary codes capable of correctingdeletions, insertions, and reversals. Soviet PhysicsDoklady 10, 8(1966), 707-710.
[15]. Li, Q. Cyberbullying in Schools. School PsychologyInternational 27, 2(2006), 157-170.
[16]. Lou, J.K., Chen, K.T., and Lei, C.L. A collusion-resistant automation scheme for social moderationsystems. IEEE Consumer Communications andNetworkingConference,2009.,(2009),571--575.
[17]. Nguyen, D. and Rose, Carolyn. Language use as areflection of socialization in online communities.Proceedings of the Workshop on Language in SocialMedia., (2011).
[18]. Sheng,V.S.,Provost,F.,andIpeirotis,P.G.Getanother label? improving data quality and data miningusing multiple, noisy labelers. Proceeding of the 14thACM SIGKDD international conference on Knowledgediscoveryand data mining-KDD '08, (2008),614.
[19]. Siersdorfer,S.,Chelaru,S.,Nejdl,W.,and San Pedro,
[20]. J. How useful are your comments?: analyzing andpredicting youtube comments and comment ratings.Proceedings of the 19th international conference onWorld wide web,(2010), 891–900.
[21]. Sood, S.O., Churchill, E., and Antin, J. Automaticidentification of personal insults on social news sites.Journal of the American Society for InformationScience andTechnology, (2011).
[22]. Spertus, E. Smokey: Automatic Recognition of HostileMessages.INPROC. IAAI, (1997), 1058--1065.
[23]. Sukumaran, A., Vezich, S., McHugh, M., and Nass, C.Normativeinfluences onthoughtful onlineparticipation. Proceedings of the 2011 annualconference on Human factors in computing systems,ACM(2011), 3401–3410.
[24]. Tetreault, J.R., Filatova, E., and Chodorow, M.Rethinking Grammatical Error Annotation andEvaluation with the Amazon Mechanical Turk.Proceedings of the NAACL HLT 2010 Fifth Workshopon Innovative Use of NLP for Building EducationalApplications, (2010), 45- 48.
[25]. Wang, J.Z., Li, J., Wiederhold, G., and Firschein, O.System for ScreeningObjectionableImages.COMPUTER COMMUNICATIONS JOURNAL 21,(1998), 1355--1360.
[26]. Yin, D., Xue, Z., Hong, L., Davison, B., Kontostathis,A.,and Edwards,L.DetectionofHarassmentonWeb2.0.ProceedingsoftheContentAnalysisintheWEB2.0(CAW2.0)WorkshopatWWW2009,(2009).
[27]. Yoon,T., Park, S.-Y., andCho, H.-G. ASmartFiltering System for Newly Coined Profanities byUsing Approximate String Alignment. Proceedings ofthe 2010 10thIEEE InternationalConferenceonComputer and Information Technology, IEEEComputer Society (2010), 643–650.