A

Analyzing and Mining Comments and Comment Ratings on the Social Web

STEFAN SIERSDORFER, L3S Research Center
SERGIU CHELARU, L3S Research Center
JOSE SAN PEDRO, Telefonica Research
ISMAIL SENGOR ALTINGOVDE, Middle East Technical University
WOLFGANG NEJDL, L3S Research Center

An analysis of the social video sharing platform YouTube and the news aggregator Yahoo! News reveals the presence of vast amounts of community feedback through comments for published videos and news stories, as well as through meta ratings for these comments. This paper presents an in-depth study of commenting and comment rating behavior on a sample of more than 10 million user comments on YouTube and Yahoo! News. In this study, comment ratings are considered first class citizens. Their dependencies with textual content, thread structure of comments, and associated content (e.g. videos and their meta data) are analyzed to obtain a comprehensive understanding of the community commenting behavior. Furthermore, this paper explores the applicability of machine learning and data mining to detect acceptance of comments by the community, comments likely to trigger discussions, controversial and polarizing content, and users exhibiting offensive commenting behavior. Results from this study have potential application in guiding the design of community oriented online discussion platforms.

Categories and Subject Descriptors: H.4 [**Information Systems Applications**]: Miscellaneous

General Terms: Algorithms, Experimentation, Measurement

Additional Key Words and Phrases: comment ratings, community feedback, youtube, yahoo! news

1. INTRODUCTION

The rapidly increasing popularity and data volume of modern Web 2.0 content sharing applications is based on their ease of operation even for unexperienced users, suitable tools for supporting collaboration and the attractiveness of shared content, e.g. images in Flickr, videos in YouTube, etc. Beyond providing access to this content, these applications offer several social mechanisms for community interaction.

One of the most prevalent mechanisms for community interaction in Web 2.0 sites is the possibility to comment on posted content and, in addition, to rate comments

A preliminary version of this paper appeared in the Proceedings of the 19th International Conference on World Wide Web (WWW '2010) [Siersdorfer et al. 2010].

This work was partially funded by the European Commission FP7 under grant agreements No. 287704 for the CUBRIK project, No. 619525 for the QualiMaster project and The Scientific and Technological Research Council of Turkey (TÜBİTAK) under the grant no. 113E065. I. S. Altingovde acknowledges the Yahoo! Faculty Research and Engagement Program.

Authors' addresses: S. Siersdorfer (corresponding author, email: siersdorfer@l3s.de), S. Chelaru, and W. Nejdl, L3S Research Center, Hannover, Germany; J. S. Pedro, Telefonica Research, Barcelona, Spain; I. S. Altingovde, Middle East Technical University, Ankara, Turkey.

DOI:http://dx.doi.org/10.1145/0000000.0000000

written by other users. Comment ratings serve the purpose of helping the community in filtering relevant opinions more efficiently. Furthermore, because negative votes are also available, comments with offensive or inappropriate content can be easily skipped. Therefore, comments and associated ratings constitute a potentially interesting data source to mine for obtaining implicit knowledge about shared content as well as community interests and interaction behavior.

The analysis of how users react to shared content and how they interact with each other is as important as the analysis of the content itself. Specifically, understanding these community dynamics has potential application to improving search and recommendation for content and comments based on these interactions. It is also the foundation for sociological and anthropological studies focused on the current trend towards digital social interactions [Kietzmann et al. 2011; Hanna et al. 2011; Susarla et al. 2012]. We have also witnessed the importance of monitoring and analyzing Web communities in the fields of journalism and politics. For instance, very recently, Obama's victory in the USA 2012 election was highly influenced by the smart use of data analysis techniques from social networks[1].

Most importantly, the study of community interaction through the commenting feature provides the basis for many technological advances targeted at improving online communities by:

(1) devising ways to encourage participation,
(2) facilitating access to relevant content and comments,
(3) and creating a safer and more appealing environment.

In this work, we provide an in-depth analysis of comments and ratings found in online community websites. The analysis focuses on providing tools for increasing user engagement in community websites by considering the three specific goals enumerated above. Although there is a large body of work that investigates and leverages comments in various social platforms (e.g. [Potthast et al. 2012; San Pedro et al. 2012; Filippova and Hall 2011; Tsagkias et al. 2010]), to the best of our knowledge, we are the first to consider comment *ratings* as first class citizens and exploit them to gain a better understanding of comments, content, and users.

We consider the analysis of two different datasets obtained from popular online communities, namely YouTube[2], an online community centered around user generated content, and Yahoo! News[3], which is centered around editorial and curated content. YouTube is the most popular video sharing site, and traffic to/from this site accounts for over 20% of the web total and 10% of the whole internet [Cheng et al. 2007], and comprises 60% of the videos watched on-line [Gill et al. 2007]. Yahoo! News is one of the leading news aggregators, and attracts a large number of users who follow news stories around the world and comment on them with over 138 million distinct yearly visitors. Both platforms provide a rich sample of comments and associated metadata for a variety of content objects from a large pool of users. Our rationale for using these two datasets is to cover different types of social media applications with different underlying incentives and commenting behavior. Information shared in these two online communities is not just different in terms of modality, but also in the particular characteristics that determine its relevance to users: videos in YouTube are commonly retrieved by specific queries issued by users, while news stories are mainly browsed in inverse chronological order in each of the predefined categories of the site. In addition to providing insights into several properties of comments and ratings, we also identify

[1]http://edition.cnn.com/2012/11/07/tech/web/obama-campaign-tech-team
[2]http://www.youtube.com
[3]http://news.yahoo.com/

Table I: Research questions addressed in this paper.

High Level question	Analysis Focus	Research Questions	Relevance
Who is commenting?	Commenter features: What can we learn about commenters by characterizing their use of language?	Can we identify troll behaviour from the history of comments of users? (Section 9)	Limiting anti-social behaviour leads to a better quality of community interaction.
What do users comment about?	Community meta-feedback as a proxy for content analysis: What can we learn about content from analyzing comments and community meta-feedback?	Does the distribution of comment ratings help identifiy characteristics of the content that generated the comments? (Section 8)	Identification of controversial shared content can be exploited as a facet in comment search.
How do users comment?	Comments and associated community feedback. We specifically consider meta-feedback (ratings) about primary community feedback (comments).	What are the intrinsic features of comments that are most liked/disliked? (Sections 3 and 4) Can we model community meta-feedback for comments? (Section 5) Can we predict comments triggering discussions or controversy? (Sections 6 and 7)	Increase the visibility of comments that can trigger community reaction, increase user participation and engagement, understand community/human behavior, identify controversial topics or objects

differences between the two datasets that occur due to behavioral differences of their corresponding communities.

We divided our work into three main types of analyses (cf. Table I), each of them focused on a different aspect of online communities:

(1) We analyze the *comments themselves* ("*How* do users comment?"). We analyze how language used and polarity of opinions in comments influence their perceived value by the rest of the community, triggers further discussion, or divides the community. We also show that machine learning models trained using comment-centric information can be leveraged to predict comments that are likely to receive high ratings in the future, initiate a discussion thread, or create controversy within the community. Promoting such comments can help to improve user satisfaction by enabling smart comment ranking methods and ultimately fuel user interaction and engagement with the underlying system. Intelligent feedback mechanisms can also benefit from this knowledge to provide users with guidelines for commenting behavior well received by and useful to other users.
(2) We analyze how the particular features of *shared content* influences community interaction and leads to polarized discussions ("*What* do users comment about?"). Detecting content that polarizes the community can be useful for providing an additional facet for retrieval and exploratory search, as is very commonly sought by a

big number of users. This can be also applied for understanding community dynamics and studying the evolution of controversy along time.

(3) We analyze individual *users* that negatively influence the interaction in the community, compromising the experience of fellow users ("*Who* is commenting?"). We study textual content of comments and their ratings for detecting *trolls*, i.e., "users who post disruptive, false or offensive comments in online communities to fool and provoke others [J.Kunegis and C.Bauckhage 2009]". Because of their organic disruptive nature, the presence of trolls can highly compromise user engagement in online communities. We study the ability of machine learning tools to detect trolls to allow for early detection and excision from the system.

2. RELATED WORK

A review of previous literature reveals a number of works that have been leveraging user comments in a wide range of different problem settings. One relevant application of user comment mining is the enhancement of retrieval mechanisms in online communities and social platforms [Potthast et al. 2012]. For instance, [Mishne and Glance 2006] investigate the impact of comments on the retrieval performance for weblogs. They find that while involving comment text in scoring does not help to improve precision, it allows for retrieving both relevant and highly discussed blog posts as an alternative to retrieving only relevant answers. In [Yee et al. 2009], the authors demonstrate the potential of comments for improving the effectiveness in a known-item retrieval scenario for YouTube. Chelaru *et al.* [2012] employ several features derived from the comments, in addition to other social signals, to improve the video retrieval effectiveness of machine-learned rankers. In [San Pedro et al. 2012] user comments are leveraged to determine the visual quality of images and to compute an aesthetic-aware re-ranking of image search results. [Agichtein et al. 2008] make use of lexical and social graph characteristics of comments and commenters in the network to find high quality content in the popular community answering system Yahoo! Answers.

Further tasks that make use of comment content include summarization of blog posts [Hu et al. 2008], prediction of video categories in YouTube [Filippova and Hall 2011], identification of political orientation in news articles based on comment sentiments [Park et al. 2011], analysis and prediction of the popularity of the commented items (e.g., [Mishne and Glance 2006; Tsagkias et al. 2010; Yano and Smith 2010]), and recommendation of related content based on commented items [Li et al. 2010; Shmueli et al. 2012]. In none of these works, comment ratings are analyzed or taken into account in the first place. In a recent study, Potthast et al. categorize retrieval related tasks for comments into two groups, namely, comment-targeting and comment-exploitation [Potthast et al. 2012]. The former group of works considers the comments themselves as retrieval targets, whereas in the latter case the retrieval targets consist of the commented items. The majority of the works that fall into the comment-targeting group essentially focused on product or movie reviews but do not address comments and comment ratings in general.

Work on sentiment classification and opinion mining such as [Pang et al. 2002; Thomas et al. 2006] deals with the problem of automatically assigning opinion values (e.g. "positive" vs. "negative" vs. "neutral") to documents or topics using various text-oriented and linguistic features. Work in this area makes also use of Senti-WordNet to improve classification performance [Denecke 2008]. However, the problem setting in these papers differs from ours as we analyze *community feedback* for comments rather than trying to predict the sentiment of the comments themselves.

There is a body of work on analyzing product reviews and postings in forums. In [Danescu-Niculescu-Mizil et al. 2009] the dependency of helpfulness of product reviews from Amazon users on the overall star rating of the product is examined and

a possible explanation model is provided. "Helpfulness" in that context is defined by Amazon's notion of how many users rated a review and how many of them found it helpful. Lu *et al.* [Lu et al. 2009] use a latent topic approach to extract rated quality aspects (corresponding to concepts such as "price" or "shipping") from comments in ebay. In [Wu and Huberman 2008] the temporal development of product ratings and their helpfulness and dependencies on factors such number of reviews or effort required (writing a review vs. just assigning a rating) are studied. The helpfulness of answers on the Yahoo! Answers site and the influence of variables such as required type of answer (e.g. factual, opinion, personal advice), topic domain of the question or "a priori effect" (i.e. Did the inquirer conduct some a priori research on the topic?) is manually analyzed in [Harper et al. 2008]. Kim *et al.* [Kim et al. 2006] rank product reviews according to their helpfulness using different textual features and meta data. However, they report their best results for a combination of information obtained from the star ratings (e.g. deviation from other ratings) provided by the authors of the reviews themselves; this information is not available for all sites, and in particular not for *comments* in YouTube and Yahoo! News. Weimer *et al.* [Weimer et al. 2007] make use of a similar idea to automatically predict the quality of posts in the software online forum Nabble.com. In comparison, our paper focuses on *community ratings for comments and discussions* rather than product ratings and reviews.

Only a few recent works focus directly on comment ratings. In [Hsu et al. 2009] a regression model is proposed for ranking comments from the social news aggregator *Digg.com* based on the community ratings. The authors studied the impact of different comment features like visibility, reputation of the comment authors, and the actual content of the comments. In [Dalal et al. 2012], the authors propose a multi-objective comment ranking strategy for 200 articles, each with 50 comments from Yahoo! News. While we also apply machine learning for comment rating prediction in YouTube (cf. Section 5), our analyses and scenarios presented here cover a much wider scope than solely predicting ratings. In [Mishra and Rastogi 2012] unsupervised, semi-supervised and active learning strategies are employed on the user-comment graph to correct the bias in comment ratings. The analysis of bias in ratings is not in the scope of our work, though we consider their findings complementary to our research directions.

Detecting abusive users (e.g., spammers, vandals, or trolls) is another topic that has recently drawn a lot of attention in the context of social and collaborative platforms such as forums or wikis: For the specific case of comments, an earlier study [Veloso et al. 2007] proposes to use associative classification to separate "good" comments from "bad" ones in order to enable automatic moderation in Slashdot, a popular technology news web site. The proposed classifier uses features based on the comments content and the social network (fans, friends, etc.) of the commenter. In another study addressing the same problem [Potthast et al. 2012], the features used for classification represent the comment quality, and include comment length, readability, frequency of vulgar terms, etc. Our approach presented here differs from these works in that we detect troll *users* instead of individual comments. In [J.Kunegis and C.Bauckhage 2009], troll users in Slashdot are detected using global and node-level social graph characteristics. In contrast, we use a bag-of-comments model for users to classify the trolls. We further show that comment ratings can be a good indicator for detecting trolls.

Works on predicting discussion threads usually aim at detecting the content items (such as news articles [Tsagkias et al. 2010; Tatar et al. 2011], tweets [Rowe et al. 2011b], or forum posts [Rowe et al. 2011a]) that are likely to attract comment replies. In [Mishne and Glance 2006], machine learning methods are used to identify disputative threads and in [De Choudhury et al. 2009], a framework is developed for characterizing the interestingness of threads based on their themes and participants. [Gómez et al. 2008] focus on the Slashdot network and repurposes h-index as an effective met-

ric of content controversy, where the number of nested replies for each comment is used as the h-index equivalent of *number of citations* for each paper. In contrast to these works, we predict the individual *comments* that are likely to attract other comments and start a discussion thread.

A complementary body of work has targeted the characterization of structural features of comment threads. In [Schuth et al. 2007], authors proposed a modeling approach to infer from a flat list of comments the hierarchical structure of posts and replies. [Gómez et al. 2012] employ generative models of growing trees to analyze the structure and evolution of discussion threads taking into account the features like popularity, novelty and bias. [Wang et al. 2012] conduct a comprehensive study of growth dynamics in conversation threads, and propose models able to explain the structural differences between popular social networks, such as Reddit, Digg and Epinions.

This paper is an extension of our previous study [Siersdorfer et al. 2010] on comments and comment ratings in YouTube. Our current submission significantly extends that work by adding a new dataset and defining three new scenarios in which we analyze and exploit comments and their ratings. Our *new* contributions in this paper can be summarized as follows.

— We extend the study by considering an additional comment corpus collected from the Yahoo! News website with the goal of identifying similarities and differences in the commenting behavior of these two communities (Section 3).
— We take advantage of the characteristics of the new corpus to perform additional studies. For instance, given that Yahoo! News provides a decomposition of comment ratings into likes and dislikes, we conduct a study of controversial comments that split the community, and further develop a prediction model for such comments (Section 7).
— We analyze discussion threads and build a machine learning model to predict *comments* that will attract replies (Section 6).
— We compare language and ratings for troll and non-troll users, and leverage the textual content of user comments for troll detection (Section 9).
— We extend sections of our previous work to include additional studies and experiments , e.g. a temporal analysis of rating behavior (cf. Section 7) and detection of controversial latent topics (cf. Section 8).

3. DATA

For this study, we gathered comments from two highly popular, community-oriented websites: YouTube and Yahoo! News.[4] YouTube is a video sharing platform where users can upload their own videos and watch, rate and comment on other users' content. At Yahoo! News users can follow news stories around the world and comment on them. Both platforms provide tools for replying other users' comments and rating them via like/dislike buttons.

Our rationale for using these two datasets is to cover different types of social media applications with different underlying incentives and commenting behavior. Information shared in these two online communities is not just different in terms of modality, but also in the particular characteristics that determine its relevance to users: videos in YouTube are commonly retrieved by specific queries issued by users, while news stories are mainly browsed in inverse chronological order in each of the predefined

[4]In section 9 we will introduce an additional dataset gathered from Slashdot (http://slashdot.org/) which will be only used in the context of troll detection.

Table II: Descriptive statistics for the YouTube and Yahoo! News corpora.

	YouTube	Yahoo! News
Mean #comments	261.75	140.94
Median #comments	13	3
Max. #comments	128,307	48,051
Stddev. #comments	2,053.84	866.43
Mean #words	8.20	15.68
Median #words	5	9
Mean #sentences	1.82	2.76
Median #sentences	1	2
Mean rating	0.61	1.39
Median rating	0	0
Stddev. rating	8.42	10.95
Max. rating	4,170	4,327
Min. rating	-1,918	-1,018

categories of the site. We are aware that the differences in the crawling methods used to collect each of the datasets produces two collections with intrinsically distinct characteristics. However, both crawling strategies aim at replicating the common retrieval interaction of users in each website and allow for comparing the particularities of comments in two widely used online communities.

3.1. Data Gathering

YouTube *collection.* We used the YouTube search engine to create this first collection by formulating textual queries. We selected our set of seed queries from Google's Zeitgeist archive from 2001 to 2007 in order to obtain a set of popular queries, similarly to our previous work [Siersdorfer et al. 2009]. We obtained a total of 756 different keywords. The top 50 results for each query were collected. We then extended this set using YouTube's "related videos" option over a sample of the already collected videos chosen uniformly at random. This scraping methodology aims at mimicking the typical user interaction (searching for videos and subsequent browsing of related/suggested videos) in YouTube. For each selected video we gathered the first 500 comments (if available), along with contextual metadata, including authors, timestamps and comment ratings (At the time of our crawl YouTube computed comment ratings by aggregating the number of likes ("thumbs up") and dislikes ("thumbs down") from users into one single value.). In addition, for each video we collected metadata such as title, tags, category, description, upload date as well as statistics provided by YouTube such as overall number of comments, views, and video rating. The complete collection had a final size of $67,290$ videos and over 6 million comments.

Yahoo! News *collection.* In order to form our second collection, we first collected all stories published in the Yahoo! News RSS feed between September and December 2011. For each story, we crawled all available comments along with their ratings (i.e., the number of likes and dislikes per comment) and replies, as well as associated meta data including the authors of comments, locations (if stated in the author profile), and timestamps. This process yielded a collection of 5.4 million comments for $27,000$ news stories.

3.2. Data Characteristics

In Table II, we provide descriptive statistics about our collections. For Yahoo! News, we observe a mean value of $\mu_{comm} = 140.94$ (median value of 3) comments per story,

whereas for YouTube the mean number of comments per video is $\mu_{comm} = 261.75$ (median value of 13). These figures reflect the actual number of comments as reported by the corresponding systems, and not according to the number of crawled comments. The difference in the average number of comments is not unexpected: as news stories are updated rapidly, they are actively accessed only for a short time. In particular, top news stories are archived for 7 days only, enforcing a natural limitation on the number of comments a story can get. On the other hand, YouTube videos feature a longer life span, allowing further comments to be added by the viewers. To give an example, the most commented news story had a total of $48,051$ comments, while the most commented YouTube video in our collection attracted $128,307$ comments.[5] In contrast, comments posted for news stories are almost twice as long as those posted for videos, both in terms of the number of words and the number of sentences[6]. A closer inspection of the datasets revealed that users commenting on a news story tend to elaborate more on concepts and opinions, whereas users in YouTube often post very short comments that simply express favor or disfavor of video clips.

We also inspected the vocabulary of all comments for each dataset, and found that YouTube and Yahoo! News comments include $702,000$ and $612,000$ terms respectively. The overlap between lexicons is about 25% ($165,000$ terms). This lexical divide is mostly due to the different topics covered in the two datasets, but also caused by the commenting behavior being organically different in both sites, which is also suggested by the differences noted above (cf. Table II). However, this should be interpreted cautiously, because both comment datasets include a high number of words with typos, abbreviations, etc.

There are also trends that are exhibited in both collections. Figure 1 shows the distribution of the number of comments per video and news story in the YouTube and Yahoo! News collections. The distributions follow the expected zipf-like pattern, characterized by having high frequencies for a relatively small number of top ranked elements and a subsequent long tail of additional low-represented elements [Cha et al. 2007].

In Table II, we also provide basic descriptive statistics about comment ratings. For the YouTube collection, we observe that ratings range from $-1,918$ to $4,170$ with a mean value of $\mu_r = 0.61$. For Yahoo! News, the ratings range from $-1,018$ to $4,327$ with a mean value of $\mu_r = 1.39$. While the minimum and maximum values for comment ratings in both datasets are in the same scale, on average, comments are rated higher in Yahoo! News than in YouTube. For a more detailed inspection, we show the distribution of comment ratings for both datasets in Figure 2. The following two main observations can be made. First, the distribution is asymmetric for positive and negative ratings, indicating that the community tends to cast more positive than negative votes. This behavior is more dominant in the Yahoo! News collection, as the mean rating score is almost twice as high as for video comments. Second, comments with rating 0 represent about 50% and 30% of the overall population for the YouTube and Yahoo! News collections, respectively, indicating that a substantial fraction of comments lack votes or are neutrally evaluated by the community.

[5]The video with the highest number of comments in our YouTube dataset is http://www.youtube.com/watch?v=lj3iNxZ8Dww ("Miss Teen USA 2007 - South Carolina answers a question"). The Yahoo! News article with the highest number of comments is http://news.yahoo.com/blogs/new-york/muslims-police-scuffle-rye-playland-over-amusement-park-123309825.html, reporting about incidents between the Muslim Community and the police in New York; these incidents were highly debated in the US.
[6]Stopwords were removed and sentences were segmented using the GATE tool available at http://gate.ac.uk/

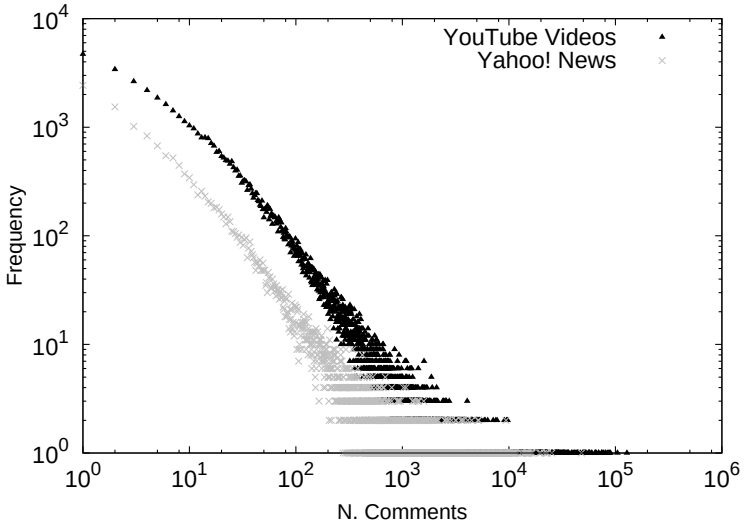


Fig. 1: Distribution of number of comments for videos in YouTube and news stories in Yahoo! News.

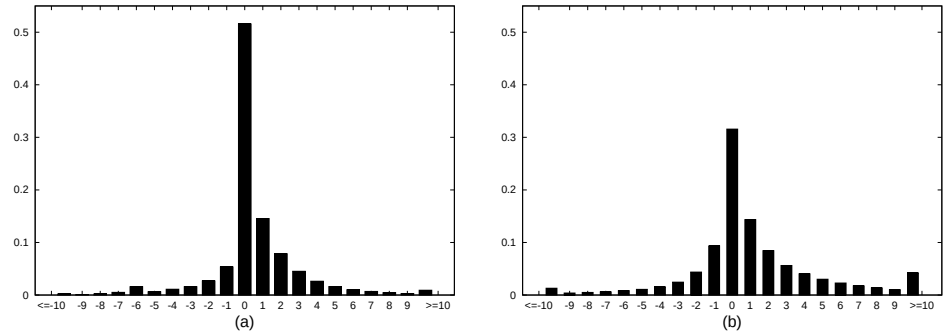


Fig. 2: Distribution of comment ratings for (a) YouTube, and (b) Yahoo! News.

3.3. Preliminary Term Analysis

The textual content of comments in Web 2.0 infrastructures can provide clues about their potential acceptance by the community. As an illustrative example we computed a ranked list of terms from a set of 100,000 comments with a rating of 5 or higher (high community acceptance) and another set of the same size containing comments with a rating of -5 or lower (low community acceptance). For tokenization of comments the java class StreamTokenizer[7] was used which considers spaces, periods, commas, and semicolons as separators. In addition, we removed stopwords[8] and applied Lucene's SnowBallAnalyzer for stemming. For ranking the resulting terms, we used the Mutual Information (MI) measure [Manning and Schuetze 1999; Yang and Pedersen 1997] from information theory which can be interpreted as a measure of how much the joint distribution of features X_i (terms in our case) deviate from a hypothetical distribution

[7]http://docs.oracle.com/javase/6/docs/api/java/io/StreamTokenizer.html
[8]http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

Table III: Top-50 terms according to their MI values for accepted comments (with high comment ratings) vs. not accepted comments (with low comment ratings).

Terms for Accepted Comments									
YouTube					Yahoo! News				
love	lt	perform	john	man	illeg	govern	famili	money	law
song	xd	nice	greatest	talent	job	politician	need	time	holder
great	britney	omg	cri	sweet	pay	doe	gun	texa	live
amaz	voic	jame	scene	inspir	border	union	feder	hope	spend
beauti	hot	movi	whitney	absolut	polit	speech	citizen	go	cheney
awesom	perfect	sexi	brilliant	allison	state	hurrican	countri	new	storm
cute	time	gorgeous	wonder	hill	friend	thing	home	taxpay	compani
favorit	miss	3	luv	ador	sad	mexico	rule	crimin	dollar
music	feel	heart	part	fantast	work	fire	bless	vacat	busi
lol	rock	rocki	scientolog	cool	campaign	immigr	don't	stay	servic

Terms for Unaccepted Comments									
YouTube					Yahoo! News				
fuck	ur	kill	die	comment	republican	tea	obama	white	parti
suck	hate	fake	bore	wtf	gop	jew	bush	liber	presid
gay	dick	idiot	crap	asshol	cain	god	hate	black	israel
shit	white	dumb	loser	horribl	jesus	racist	rich	america	gay
bitch	fat	bad	hell	whore	bagger	christian	2012	conserv	teabagg
stupid	black	guy	peopl	Im	wing	vote	fox	right	american
ass	fag	obama	shut	lame	like	fact	world	bibl	protest
nigger	faggot	de	worst	racist	lol	truth	look	zionist	win
ugli	jew	cunt	fuckin	hey	lie	class	christ	jewish	kill
dont	retard	pussi	cock	read	democrat	earth	nazi	herman	die

in which features and categories ("high community acceptance" and "low community acceptance") are independent from each other.

Table III shows the top-50 (stemmed) terms[9] extracted for each category. Obviously many of the "accepted" comments in the YouTube collection contain terms expressing sympathy or commendation (*love*, *fantast*, *greatest*, *perfect*). "Unaccepted" comments, on the other hand, often contain swear words (*retard*, *idiot*) and negative adjectives (*ugli*, *dumb*); this indicates that offensive comments are, in general, not promoted by the community. We applied the same term analysis procedure on the Yahoo! News collection. The difference between terms from the "accepted" and "unaccepted" categories is still visible but not as significant as for YouTube. Yahoo! News is more sensitive to the language used by users and it enforces stricter policies concerning insults and hate phrases [10], which makes the content in accepted and unaccepted comments lexically more similar.

3.4. Summary and Lessons Learned

This section introduced the two data sources used for our study of commenting characteristics in online communities: YouTube and Yahoo! News. YouTube videos tend to attract a larger number of comments than Yahoo! News, explained by the shorter life expectancy of news stories compared to YouTube video clips. However, Yahoo! News comments are twice as long as YouTube's on average, which can be explained by the nature of the commented content: news stories tend to produce comments where users describe their opinions while videos attract comments of simple praise or rejection.

[9]Note that some of the terms seem to emerge from the use of emoticons or similar symbols. For instance, the sequence "<3" is used to represent a heart symbol. On the other hand, "de" might be part of a URL (".de" for Germany).

[10]http://help.yahoo.com/kb/index?locale=en_US&page=content&y=PROD_NEWS&id=SLN2292

The distribution of comment ratings shows a clear tendency towards positive ratings in both datasets (Figure 2). This similarity emerges despite the significant differences previously described, which hints at a natural users bias towards positivity in rating assignments. A study of the most discriminative terms for comments being rated as positive or negative revealed clear lexical patterns that support the attempt to use automatic classification methods for its prediction (Table III), which we exploit in Section 5. The sentiment revealed in our preliminary term analysis is systematically analyzed in Section 4.

4. SENTIMENT ANALYSIS OF RATED COMMENTS

In this section, we make use of the publicly available SentiWordNet thesaurus to study the connection between the sentiment features of comments and the ratings they get. We aim at understanding how *the way users express opinions* affects comment approval from the rest of the community, regardless of the actual opinion stated.

4.1. Preliminaries: SentiWordNet

SentiWordNet [Esuli and Sebastiani 2006] is a lexical resource built on top of WordNet. WordNet [Fellbaum 1998] is a thesaurus containing textual descriptions of terms and relationships between terms (examples are hypernyms: "car" is a subconcept of "vehicle" or synonyms: "car" describes the same concept as "automobile"). WordNet distinguishes between different part-of-speech types (verb, noun, adjective, etc.) A *synset* in WordNet comprises all terms referring to the same concept (e.g. $\{car, automobile\}$). In SentiWordNet a triple of three *senti values* (pos, neg, obj) (corresponding to positive, negative, or rather neutral sentiment flavor of a word respectively) are assigned to each WordNet synset (and, thus, to each term in the synset). The sentivalues are in the range of $[0, 1]$ and sum up to 1 for each triple. For instance $(pos, neg, obj) = (0.875, 0.0, 0.125)$ for the term "good" or $(0.25, 0.375, 0.375)$ for the term "ill". Sentivalues were partly created by human assessors and partly automatically assigned using an ensemble of different classifiers (see [Esuli 2008] for an evaluation of these methods).

4.2. Sentiment Analysis of Ratings

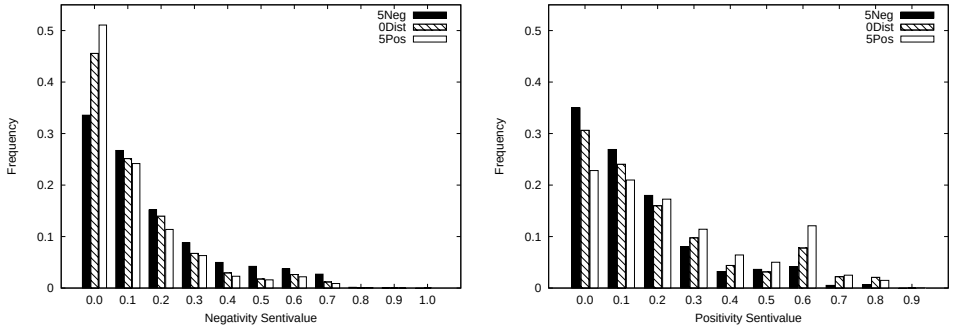
We now describe our statistical comparison of the influence of sentiment on comment ratings. In our experiments, we assigned a sentivalue to each comment by computing the averages for pos and neg over all words in the comment that have an entry in SentiWordNet. We restrict our analysis to adjectives, as we observed the highest accuracy in SentiWordNet for these.

Our intuition is that the choice of terms used to compose a comment may provoke strong reactions of approval or denial in the community, and therefore determine the final rating score. For instance, comments with a high proportion of offensive terms would tend to receive more negative ratings. We used comment-wise sentivalues, computed as explained above, to study the presence of sentiments in comments according to their rating.

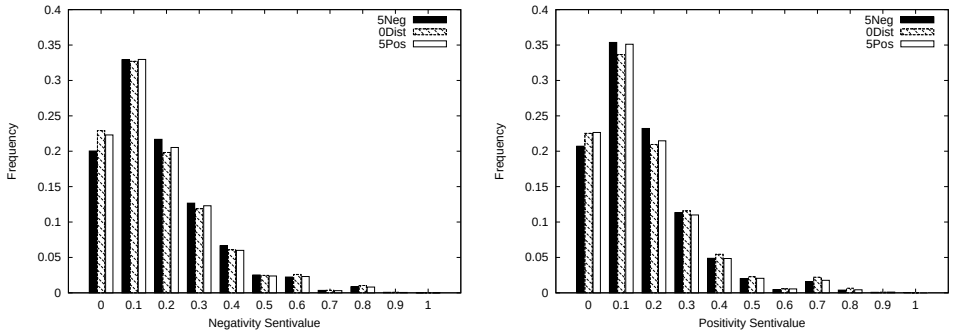
To this end, we first subdivided the data set into three disjoint partitions:

— **5Neg**: The set of comments with rating score r less or equal to -5, $r \leq -5$.
— **0Dist**: The set of comments with rating score equal to 0, $r = 0$.
— **5Pos**: The set of comments with rating score greater or equal to 5, $r \geq 5$.

We then analyzed the dependent sentiment variables "positivity" and "negativity" for each different partition. Detailed comparison histograms for these sentiments are shown in Figure 3. This figure shows a clearly different behavior in YouTube and Yahoo! News. In the case of YouTube, the results follow our intuition: negatively rated comments (**5Neg**) tend to contain more negative sentiment terms than positively rated



(a) YouTube



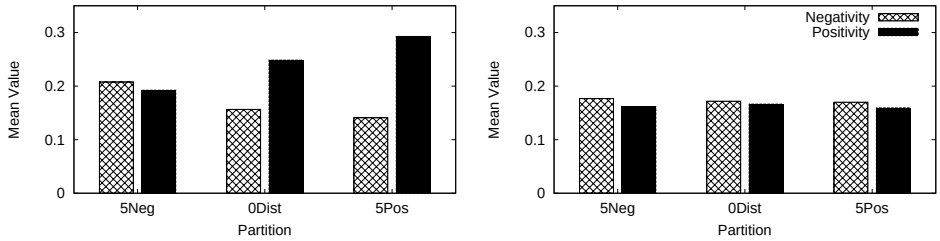
(b) Yahoo! News

Fig. 3: Distribution of comment negativity, and positivity for (a) YouTube and (b) Yahoo! News

comments (**5Pos**). This is reflected by a lower frequency of sentivalues at negativity level 0.0 along with consistently higher frequencies at negativity levels ≥ 0.1. Similarly, positively rated comments tend to contain more positive sentiment terms. We also observe that comments with rating score equal to 0 (**0Dist**) have sentivalues in between, in consonance with the initial intuition.

In the case of the Yahoo! News dataset, the observed pattern is substantially different. The distribution of sentivalues, both positive and negative, does not have a clear dependency with respect to the considered partition. That is, comment ratings are not as influenced by the sentiment orientation of the words contained in them. This result is in consonance with the observations from Section 3. Comments in Yahoo! News are subject to stricter policies which reduces the occurrence of offensive terms. In addition, it is expected that well written comments could attract negative ratings just because of the diversity of opinions in the matters normally covered in the news. As a consequence ratings depend less on the sentiment orientation of comments for that platform.

We conducted tests to examine the statistical difference of the average senti-values (both positive and negative distributions) across the three groups defined (**5Neg**, **0Dist**, **5Pos**) in both datasets. To this end, we selected a random sample of $5,000$ comments for which sentiment values were available in SentiWordNet. The analysis of variance (ANOVA) test for each of these 4 conditions (2 sentiment orientations $\times$ 2 datasets) systematically resulted in a strong support of the significance of the differ-



(a) YouTube (b) Yahoo! News

Fig. 4: Comparison of mean senti-values for comments with different kinds of community ratings in (a) YouTube and (b) Yahoo! News.

ence (p-value < 0.01) across the three groups. Figure 4 shows the difference of mean values for negativity and positivity, revealing that negative sentivalues are predominant in negatively rated comments, whereas positive sentivalues are predominant in positively rated comments. This difference, however, is clearly more noticeable in the YouTube dataset in consonance with previously reported results.

4.3. Summary and Lessons Learned

This section introduces the dimension of sentiment to the analysis of comment texts. Figure 4 shows a noticeable positive correlation between the overall sentiment of terms in comments and ratings, which we showed to be statistically significant. This result provides further support for the viability of using automatic classification methods to predict the expected community acceptance of comments (Section 5).

5. PREDICTING COMMENT RATINGS

In this section we study methods for predicting comment ratings using the textual features of comments. Such methods have direct application towards smart comment ranking and filtering, and have the potential to be used for providing automatic real time feedback to writers about the acceptability of their comments. These applications can improve online community usability and could lead to higher user engagement with the platform. Our term- and SentiWordNet-based analyses described the previous sections indicate the discriminative character of terms occurring in comments with respect to comment ratings.

5.1. Experimental Setup for Classification

We used machine learning and term-based representations of comments to automatically categorize comments as likely to obtain a high overall rating or not. In order to classify comments into categories "accepted by the community" (high comment rating) or "not accepted" (low comment rating), we employed linear support vector machines (SVMs) as they have been shown to perform well for various classification tasks (see, e.g., [Dumais et al. 1998; Joachims 1998]). Comments labeled as "accepted" or "not accepted" are used to train the classification model. The feature vector of a comment was constructed using the the frequencies of the terms occurring in the comment, normalized by the number of terms in the comment. We removed stopwords[11] and terms occurring just once in the corpus, and applied Lucene's SnowBallAnalyzer for stem-

[11] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

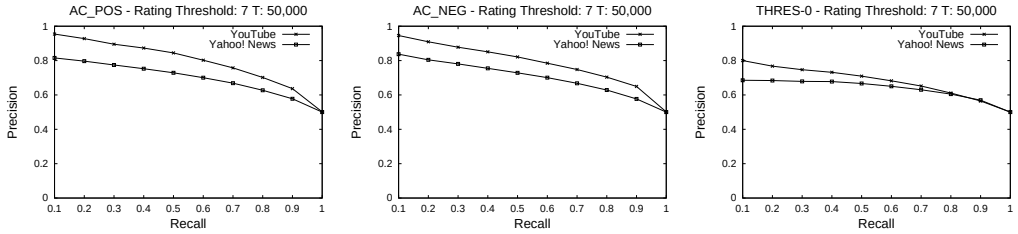


Fig. 5: Precision-recall curves for comment rating prediction.

Table IV: Comment rating classification: BEPs for different training set sizes T and different rating thresholds.

	YouTube			Yahoo! News		
T	Rating ≥ 2	Rating ≥ 5	Rating ≥ 7	Rating ≥ 2	Rating ≥ 5	Rating ≥ 7
			AC_POS			
10,000	0.659	0.696	0.721	0.577	0.624	0.649
50,000	0.679	0.715	**0.738**	0.586	0.654	**0.676**
200,000	0.693	-	-	0.597	0.668	-
			AC_NEG			
10,000	0.659	0.693	0.721	0.574	0.628	0.646
50,000	0.678	0.714	**0.734**	0.588	0.652	**0.676**
200,000	0.691	-	-	0.604	0.668	-
			THRES-0			
10,000	0.595	0.620	0.640	0.566	0.609	0.620
50,000	0.605	0.642	0.663	0.577	0.628	0.642
200,000	0.621	0.656	**0.671**	0.618	0.642	**0.656**

ming. We used the LIBSVM [Chang and Lin 2011] implementation of support vector machines using a linear kernel and cost parameter C=0.1.

How can we obtain sufficiently large training sets of "accepted" or "not accepted" comments? We are aware that the concept is highly subjective and problematic. However, the amount of community feedback in YouTube results in large annotated comment sets which can help to average out noise in various forms and, thus, reflects to a certain degree the "democratic" view of a community. To this end we considered distinct thresholds for the minimum comment rating for comments. Formally, we obtain a set $\{(\vec{c_1}, l_1), \ldots (\vec{c_n}, l_n)\}$ of comment vectors $\vec{c_i}$ labeled by l_i with $l_i = 1$ if the rating lies above a threshold ("positive" examples), $l_i = -1$ if the rating is below a certain threshold ("negative" examples).

We performed different series of binary classification experiments of YouTube comments into the classes "accepted" and "not accepted" as introduced in the previous subsection. For our classification experiments, we considered different levels of restrictiveness for these classes. Specifically, we considered distinct thresholds for the minimum and maximum ratings (above/below +2/-2, +5/-5 and +7/-7) for comments to be considered as "accepted" or "not accepted" by the community.

We also considered different amounts of randomly chosen "accepted" training comments (T = 10,000, 50,000, 200,000) as positive examples and the same amount of randomly chosen "unaccepted" comments as negative samples (where that number of

training comments and at least 1,000 test comments were available for each of the two classes). For testing the models based on these training sets we used the disjoint sets of remaining "accepted" comments with same minimum rating and a randomly selected disjoint subset of negative samples of the same size. We performed a similar experiment by considering "unaccepted" comments as positive and "accepted" ones as negative, thus, testing the recognition of "bad" comments. We also considered the scenario of discriminating comments with a high absolute rating (either positive or negative) against unrated comments (rating = 0). The three scenarios are labeled **AC_POS**, **AC_NEG**, and **THRES-0** respectively.

5.2. Results

Our quality measures are the precision-recall curves as well as the precision-recall break-even points (BEPs) for these curves (i.e. precision/recall at the point where precision equals recall, which is also equal to the F1 measure, the harmonic mean of precision and recall in that case). The results for the BEP values are shown in Table IV. The detailed precision-recall curves for the example case of T=50,000 training comments class and thresholds +7/-7 for "accepted"/ "unaccepted" comments are shown in Figure 5 for YouTube and Yahoo! News. The main observations are:

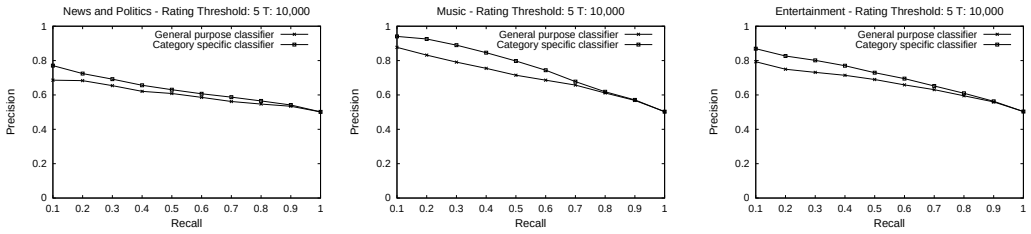
— All three types of classifiers provide good performance for the YouTube dataset. For instance, the configuration with T=50,000 positive/negative training comments and thresholds +7/-7 for the scenario **AC_POS** leads to a BEP of 0.738. Consistently, similar observations can be made for all examined configurations.
— Trading recall against precision for YouTube leads to applicable results. For instance, we obtain prec=0.872 for recall=0.4, and prec=0.954 for recall=0.1 for **AC_POS**; this is useful for finding candidates for interesting comments in large comment sets.
— Classifiers perform worse for Yahoo! News; with T=50,000 positive/negative training comments and thresholds +7/-7 we obtain a BEP of 0.676 for predicting positively rated comments. This is expected as our discriminate term and sentiment analyses described in previous sections revealed less clear patterns for that dataset. However, trading precision for recall can help again to obtain more applicable results (prec=0.815 for recall=0.1).
— Classification results tend to improve, as expected, with increasing number of training comments. Furthermore, classification performance increases with higher thresholds for community ratings for which a comment is considered as "accepted".

Overall our results confirm our intuition that there exist discriminative terms which depend on the comment ratings (see Table III in Section 3) and which enable us to train meaningful classification models.

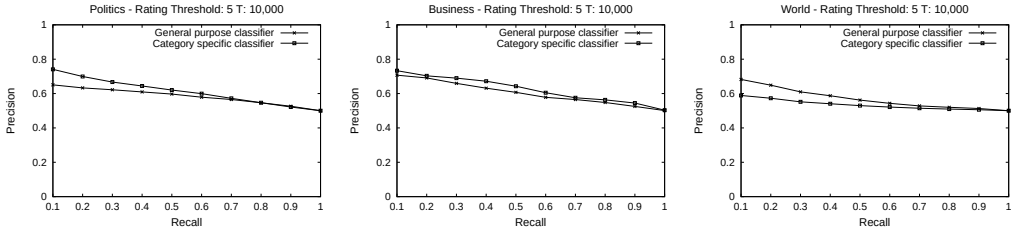
5.3. Category-Specific Rating Prediction

We also studied whether content, specifically the topic-related categories of videos and news articles, have an influence on comment rating behavior, and whether this information can be leveraged to improve classification performance for rating prediction. To this end, we conducted category specific classification experiments, using comments on videos of the same YouTube/Yahoo! News categories for testing and training. We compared the performance of these classifiers with "general purpose" classifiers using training sets randomly chosen across all categories as in the experiments described above.

Both YouTube and Yahoo! News provide category information for their videos. For YouTube we selected videos from the categories "Music", "Entertainment", and "News and Politics"; for Yahoo! News we selected "Business", "Politics", and "World". For each experiment we randomly chose 10,000 "accepted" and the same number of unaccepted



(a) YouTube



(b) Yahoo! News

Fig. 6: Precision-recall curves for category-specific rating prediction experiments.

"unaccepted" documents using a threshold of +5/-5 respectively. For each of the described categories from Yahoo! News and YouTube we selected test comments to be from the same category as the training comments, and test set sizes equal to training set sizes. We then compared the performance with that of a classifier trained on comments randomly chosen across categories. Feature vector construction, machine learning algorithm, and parameter settings were the same as in previous experiments described in this section.

Figure 6 shows the resulting precision-recall curves. We observe that the category-specific classifiers consistently outperform the "general purpose" classifiers for all tested categories both for YouTube and Yahoo! News. For instance, for the "Entertainment" category in YouTube the performance boost in the recall range of 0.1 up to 0.4 is more than 5%. Note that, in order to obtain consistent numbers of training comments on a per category level we had to restrict trainings set sizes in this experiment, resulting in a decrease of absolute performance values in comparison to our previous experiments.

5.4. Summary and Lessons Learned

In this section we considered the problem setting of inferring comment ratings using the textual content of comments. The results indicate that the average accuracy is relatively high, with YouTube outperforming Yahoo! News as a result of the more discriminative terms and more explicit sentiment present on the former (cf. Table IV). For practical purposes, precision can be traded for recall to find very accurately potential candidates for very high and low ratings; building specialized classifiers for categories in YouTube and Yahoo! News can help to further improve the performance (cf. Figure 6). Such models have direct application to comment ranking and for filtering and promotion purposes. An additional application lies in the field of intelligent user interfaces, where rating prediction could be performed in realtime to give users feedback about the acceptability of their comments as they write.

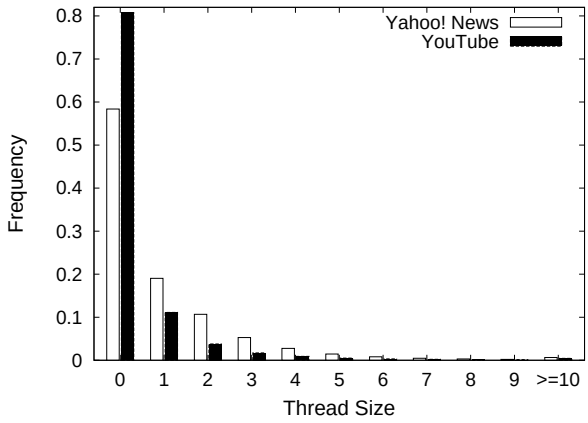


Fig. 7: Distribution of thread sizes (number of replies) for the YouTube and Yahoo! News corpora.

6. DISCUSSION THREADS AND REPLIES

Besides comment ratings, another important community feedback mechanism in many Social Web environments is the option of posting replies to comments. In this section we compare the distribution of replies for YouTube and Yahoo! News, study the relationship between comment ratings and replies, and apply machine learning to identify comments likely to receive replies. Predicting and promoting comments that are likely to trigger an online discussion can help to increase user participation and engagement within online collaborative platforms.

6.1. Distribution of Replies and Ratings

In this section we study how prominent are replies in these communities as well as what is the connection between discussion threads and ratings.

Replies in YouTube and Yahoo! News. Table V shows the basic frequency statistics for comments that received one or more replies in YouTube and Yahoo! News as well as statistics for the corresponding thread sizes. In the remainder of the paper we will

Table V: Basic statistics for seed and reply comments in the YouTube and Yahoo! News corpora.

	YouTube			Yahoo! News		
	Unreplied	Seeds	Replies	Unreplied	Seeds	Replies
Amount	3,470,413 (56.4 %)	827,603 (13.5 %)	1,851,849 (30.1 %)	1,599,228 (29.2%)	1,139,833 (20.9 %)	2,739,426 (49.9 %)
Avg. #words	7.03	11.11	9.69	16.50	21.52	12.75
Median #words	4	6	6	10	13	8
Stdev. #words	8	9.84	10.43	24.54	29.06	17.74
Avg. #sentences	1.68	2.12	2.01	3.06	3.39	2.47
Median #sentences	1	1	1	2	2	2
Avg. rating	0.90	0.27	0.19	2.24	2.60	0.38
Median rating	0	0	0	1	1	0
Stdev. rating	9.04	7.83	7.38	4.92	22.81	2.59
Min. rating	-710	-1,918	-445	-66	-1,018	-210
Max. rating	3,807	2,693	4,170	722	4,327	238

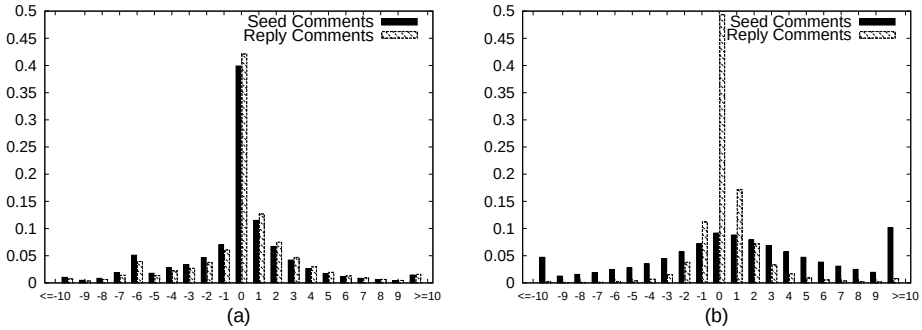


Fig. 8: Distribution of ratings for seed and reply comments in (a) the YouTube dataset and (b) the Yahoo! News dataset.

refer to comments that received at least one reply as *seed comments*. Likewise, we will refer to comments that are *not* replies to other comments as *main comments*.

We observe that Yahoo! News contains a higher proportion of seed comments among main comments than YouTube. Furthermore, almost 50% of all comments in our Yahoo! News collection are replies to other comments, in comparison to just 30% for the YouTube collection (in consonance with the 23.4% reported in [Thelwall et al. 2012] for a more recent sample of YouTube). We also observe that both seed and reply comments tend to be longer (i.e. contain more sentences and words) in Yahoo! News. These results indicate that Yahoo! News users are more likely to engage in discussions, triggered by the specific characteristics of news stories (e.g. novelty, controversy).

Figure 7 shows the distribution of the number of comments without replies and seed comments with respect to the number of replies (we define comments without replies as threads with size 0). Similar to the larger number of comment ratings already observed for the Yahoo! News collection in Section 3 (cf. Figure 2), the number of replied comments in that collection is higher at all levels of thread size. We might expect that this is due to the wider variety of content in YouTube in comparison to Yahoo! News where political topics are predominant. However, an analysis of reply behavior in YouTube restricted to videos from the category *news & politics* did not reveal any significant differences with respect to the results found for the complete YouTube collection, hinting at a homogeneous reply behavior across categories in this community. This result provides additional support that the media types (i.e. videos vs. news stories) play a central role and that the particularities of users in each community are the main responsible for these different usage patterns.

Threads and Ratings. Figure 8 shows the distribution of the average rating of seed comments and replies for both datasets. The distribution of seed comments for Yahoo! News shows longer tails for both positively and negatively rated comments, whereas ratings in YouTube concentrate around zero, following the trend shown in Figure 2. An interesting artifact to be noticed in this figure is the peak exhibited by *replies* with a rating value of zero in Yahoo! News. The most likely explanation that we found for this behavior is that the Yahoo! News web interface does not show comment replies by default. An explicit user action (clicking on a link below the seed comment) is required so replies are shown and can be rated. As a result, users are less likely to see replies and rate them.

We also studied the dependency of ratings for seed comments and the length of the corresponding discussion threads. In Figure 9 we see that for Yahoo! News the average rating of seed comments grows with increasing thread size. A possible explanation for

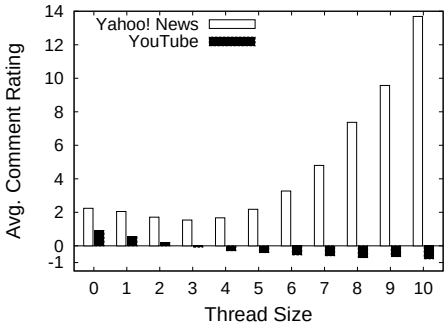


Fig. 9: Average comment rating of seed comments in Yahoo! News and YouTube with respect to thread size.

this is that comments initiating larger threads of replies have the potential of drawing more attention from the community, which helps attracting more comment ratings. Note that in general the distribution of comment ratings is skewed towards positivity in Yahoo! News (cf. Figure 8). Interestingly, for YouTube we observe a decrease of ratings, and even a trend towards negative ratings, with increasing thread size; manual inspection (conducted by one of the authors of this article) of a random sample of 300 threads containing 20 or more comments confirmed that longer discussions in YouTube often tend to be rude or "flame wars" (82.6% of the threads in our sample with a 95% confidence interval of $\pm 4\%$). More specifically, we refer to flame wars as discussions/interactions characterized by "words of profanity, obscenity, and insults that inflict harm to a person or an organization" as discussed in literature in Social Sciences [Alonzo and Aiken 2004].

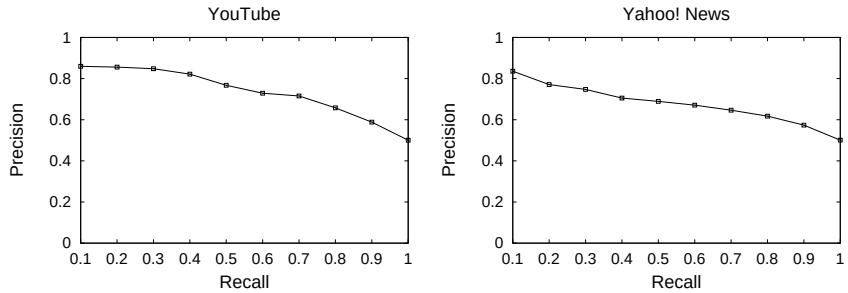
Note that, so far, we only considered the overall rating of the comments. The Yahoo! News data comprises additional information on how ratings of individual comments decompose into positive and negative votes ("likes" and "dislikes"). Section 7 will be dedicated to this topic and we will revisit discussion threads in that context.

6.2. Predicting the Responsivity on Comments

In order to study if information obtained from the textual content of comments can be used to predict seed comments (i.e., comments that receive replies and start a discussion thread), we performed different series of binary classifications of YouTube and Yahoo! News comments into the classes "Seed" and "Unreplied". Here we considered different levels of restrictiveness for these classes. Specifically, we considered distinct thresholds R for the minimum number of received replies (2,5,7, and 9) for comments to be considered as "Seeds"; comments with no replies were considered as "Unreplied". Our rationale for studying different reply thresholds was to explore how the amount of replies can influence the classifier performance. We also considered different amounts of randomly chosen "Seed" training comments (T = 5,000, 10,000, 30,000, 100,000) as positive examples and the same amount of randomly chosen "Unreplied" comments as negative samples (where that number of training comments were available for each of the two classes). We tested the models on the disjoint sets of remaining seed comments (with different thresholds R for the minimum number of received replies) and a randomly selected disjoint subset of "unreplied" comments of the same size. The sizes of the test sets varied for YouTube from 26,170 (T=10,000, R $\geq$ 9) to 680,726 (T=5,000, R $\geq$ 2) and for Yahoo! News from 26,080 (T=10,000, R $\geq$ 10k) to 1,226,248 (T=5,000, R $\geq$ 2).

Table VI: BEPs for classification of seed comments vs. comments without replies.

T	YouTube				Yahoo! News			
	R≥2	R≥5	R ≥ 7	R ≥ 9	R≥2	R≥5	R ≥ 7	R ≥ 9
5,000	0.617	0.636	0.681	0.692	0.567	0.620	0.632	0.636
10,000	0.635	0.654	0.690	0.712	0.579	0.621	0.644	0.655
30,000	0.637	0.678	-	-	0.588	0.633	-	-
100,000	0.649	-	-	-	0.591	-	-	-



(a) (b)

Fig. 10: Precision-recall curves for predicting replied comments for (a) the YouTube and (b) the Yahoo! News dataset.

We used LIBSVM with linear kernel and default parameterization. Feature vectors for comments were constructed as described in Section 5.

The resulting values of the precision-recall break-even-point (BEP) are shown in Table VI. We observe that the performance improves with increasing reply threshold R; this is expected as comments with more replies are more prominent representatives of comments triggering discussions. Classification performances are comparable for both Social Web environments studied.

The detailed precision-recall curves for the best-performing setting (i.e. $R \geq 9$) are shown in Figure 10. Due to the difficulty of the task, reply prediction is not feasible for high-recall scenarios. However, trading recall against precision leads to applicable results. For instance, given a recall of 0.1, we obtain a precision of 0.859 for the YouTube dataset and a precision of 0.836 for the Yahoo! News data set. This is useful for automatically identifying at least part of the most likely candidates for comments triggering discussions, and can still help to mine a substantial amount of interesting seed comments from large datasets.

6.3. Summary and Lessons Learned

In this section we studied replies as a new type of feedback mechanism. Firstly, we examined the number of comments receiving replies, the distribution of thread sizes, and the number of comments receiving replies (Table V, Figure 7). We observed that replies play a considerable role in both platforms; YouTube users, however, are less prone to engage in elaborate discussions despite considering highly controversial content categories, such as News and Politics. From a social computing point of view, this result leads to a number of relevant research questions. For instance: Is this different behavior caused by the different media types (video vs. text)? Does the media type have

an influence on the education level of the user communities attracted to the platform, leading to this behavior differences?

Secondly, we examined the connection between replies and comment ratings (Figures 8 and 9). Interestingly, for Yahoo! News comment ratings tend to increase with the thread size of the discussions triggered by comments, whereas for YouTube we can observe a contrary behavior due to more rude discussions and flame wars. On the other hand, the high ratings for comments leading to longer discussions in Yahoo! News indicate that these comments are appreciated by the community and lead to user engagement. Another interesting observation is that, in comparison to YouTube, for Yahoo! News there is a large number of replies without ratings. The likely reason for this is that, by default, in Yahoo! News replies are hidden from the user; here, user engagement might be further improved by showing these comments more prominently.

Finally, we evaluated the predictive performance of textual features for replies (Table VI, Figure 10). Although, due to the difficulty of the task, this turns out to be unfeasible for high-recall scenarios, trading recall against precision leads to applicable results, and can help identifying likely candidates for comments triggering discussions.

7. CONTROVERSIAL COMMENTS

So far we have explored comment ratings as a single aggregate value for the community acceptance of a given comment. However, in many Social Web environments, the overall rating score can be decomposed into a number of "likes" and "dislikes" (with overall rating score = #likes - #dislikes). This can reveal additional information about the community perception. For instance, consider a comment receiving 10 "likes" and 0 "dislikes" versus a comment receiving 100 "likes" and 90 "dislikes". Although the overall rating is the same for both comments (i.e. +10), the content of the latter comment is likely to be more controversial. More generally, in this section we study *controversial comments* which attract a more balanced number of positive and negative votes versus rather *non-controversial comments* where either the positive or the negative votes are dominant.

Table VII shows some hand-picked comments from our Yahoo! News set to illustrate both classes of comments. For instance, those comments that are supporting/criticizing either one of the democrat or republican leaders in the US are equally liked and disliked. In contrast, a general criticism about politicians is liked by almost everybody, and praising Bin Laden is disliked by almost all of the raters.

The concept of controversy is highly subjective. Controversial issues are commonly agreed to be those that trigger "disagreement between individuals" [Harwood and Hahn 1990] or "cannot be solved to everyone's satisfaction" [Kuypers 2002]. However, it is hard to achieve absolute agreement about a precise specification of that concept, leading to some fuzziness in definitions. In this article, we model agreement/disagreement between individuals by looking at the *rating divergence* for a given comment. Alternative ways of identifying controversy could also be used. In particular, we believe that metrics associated to thread size or the content of comment replies could lead to additional interesting models of controversy. Such alternative metrics might be useful in other application contexts (e.g. for analyzing online environments with no support for comment ratings). Furthermore, they might be helpful for analyzing different aspects of controversial comments (e.g. controversy strength based on sentiment analysis) and the community itself (e.g. the tendency to reply in contrast to rate controversial comments). This section presents a first step towards the study of controversial comments and is based on rating divergence; the analysis of further notions of controversy is out of the scope of this article and left as future work.

We first provide an analysis of controversial comments covering various aspects, and then focus on developing models for automatically detecting controversial comments.

Table VII: Examples of comments belonging to the categories *"controversial"* and *"non-controversial"*.

Likes	Dislikes	Text
		Controversial Comments
24	16	Why do the republicants hate working class America so much? OBAMA 2012
32	32	Bush....gasoline $1.87/gal Obama...gasoline $3.47/gal
16	20	The Tea Party hates two things. 1. Being called racist. 2. Black people
10	15	For some reason, a lot of you thing that rich people pay NO taxes? They pay taxes even though 50% of Americans do not. What Obama wants to do is RAISE their taxes. That's not fair. Let's make sure everyone pays taxes and politicians use tax money in a sensible way before we raise taxes on a few.
22	27	Sounds a whole lot like how scientists are treated who don't believe man-made global warming alarmists.
		Non-controversial Comments
118	2	we have the best politicians.......THAT MONEY CAN BUY!!
21	0	Politicians should be required to wear the logos of their corporate sponsors like race car drivers do... there would be a lot less confusion about who they're actually representing.
34	1	Washington DC has over 41000 lobbyist sorry guys business as usual
104	9	Occupy Washington DC !
2	32	Osama bin Ladin will forever be remembered as the man who brought America to it's knees. 9/11 was a blessing
117	13	One day I hope everyone who lost someone in this disaster can be at rest. Im so sorry for the ones who still hurt and ask GOD to put u at rest just knowing they r with him now and SAFE.GOD BLESS AMERICIA.

Retrieving such comments can be especially interesting for opinion researchers and journalists as it allows them to identify, in advance, comments and topics that divide the community. Note that our discussion is restricted to Yahoo! News comments because YouTube did not provide separate numbers for "likes" and "dislikes" at the time of crawling (in 2009).[12]

7.1. Distribution of "likes" and "dislikes"

We first want to study how the different proportions of "likes" and "dislikes" are distributed in Yahoo! News. For a comment c containing at least one rating, we define the *comment approval ratio* (Φ) as $\Phi(c) = \frac{l_c}{l_c+d_c}$, where l_c (d_c) represents the number of likes (dislikes) for comment c. A ratio close to 0 means that the comment is totally rejected by the community and 1 indicates complete approval/acceptance. In contrast, $\Phi(c)$ values around 0.5 correspond to *controversial comments* that received a balanced

[12]For a later crawl of YouTube we found that just a very small number of comments (about 2.5k out of 36 million) received a combination of a considerable number of likes and dislikes at the same time. This might be due to the nature of the (constantly changing) interface that was hiding comments with a certain amount of negative ratings.

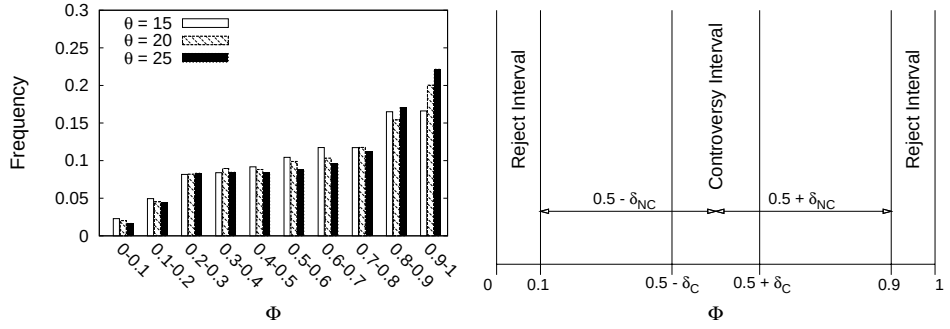


Fig. 11: (a) Distribution of number of comments per comment approval (Φ) intervals for distinct thresholds θ for the number of received ratings. (b) Controversy interval vs. accepted (positive) and not accepted (negative) intervals.

number of "likes" and "dislikes". Since comments with a higher number of ratings can provide stronger evidence for the users' opinions in comparison to those with only few ratings, we define a threshold θ for the total number of likes and dislikes received by a comment. For our study we chose θ= 15, 20, and 25.

Figure 11a shows the distribution of Yahoo! News comments with respect to their Φ values. We observe that the number of comments increases with higher values of Φ, which is expected given our earlier findings that comment ratings are skewed towards positivity (cf. Section 3). In particular, while highly-disliked comments (the first bin in the plot) constitute less than 5%, highly-liked comments (the last bin) constitute more than 15% of all comments in a given set. The positivity in ratings becomes even more dominant if we focus on comments with a larger number of ratings (i.e. higher values of θ). Regardless of the θ threshold, comments with $\Phi \in [0.4, 0.6]$ (i.e. around 0.5) add up to 20% of all comments. This shows that the number of comments causing controversy among users is relatively large.

As depicted in Figure 11b, based on the comment approval ratio, Φ, we more formally define a comment c as controversial if $0.5 - \delta_C \leq \Phi(c) \leq 0.5 + \delta_C$ where $\delta_C \in [0, 0.5]$ defines the boundaries of the *controversy interval*. Analogously, we define non-controversial comments as comments c with $0.5 - \delta_{NC} \leq \Phi(c) \leq 0.5 + \delta_{NC}$ where δ_{NC} specifies the required distance of a comment's Φ value from 0.5. In the following, unless stated otherwise, we set δ_C equal to 0.1, i.e., we consider comments for which Φ values fall into the range $[0.4, 0.6]$ as controversial comments. For the non-controversial comments, we study δ_{NC} values of 0.2, 0.3, and 0.4, with larger values of δ_{NC} corresponding to more restrictive thresholds for considering comments as non-controversial.

Figure 12 shows the Zipf-like distribution of controversial comments across the news stories for all stories with at least 1 controversial comment (absolute values, $\delta_C = 0.1$). Overall, we observe a considerable amount of comments matching our definition of controversial comments. For instance, for $\theta = 15$, around 15% percent of all stories contain at least one controversial comment posted for them.

7.2. Term Analysis

In order to examine the differences in language and vocabulary usage between controversial and non-controversial comments we conducted a discriminative term analysis. We set δ_C to 0.1, δ_{NC} to 0.4 and θ to 25, and, using Mutual Information (cf. Section 3),

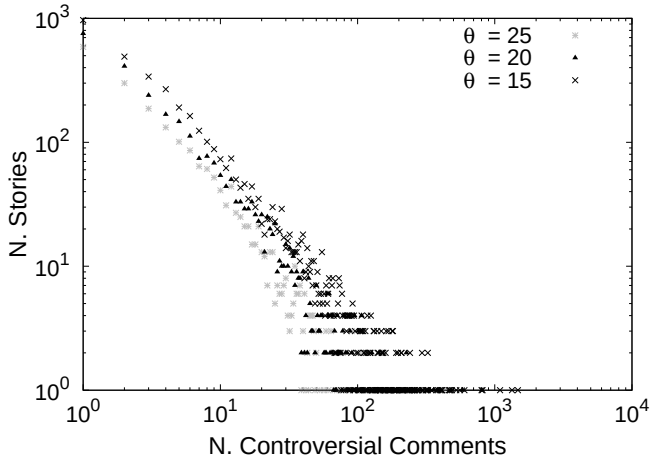


Fig. 12: Distribution of controversial comments for distinct thresholds θ for the number of received comment ratings.

we computed a ranked list of stemmed terms for approx. 6,000 comments from each of the two classes. Table VIII shows the top-20 stemmed terms extracted for each set. Many of the controversial comments contain terms related to political parties/entities involved in US presidential elections (*obama*, *republican*, *democrat*, *bush*) or terms expressing strong emotions (*believe*, *hate*). We conducted a manual inspection of comments and found that the latter type of terms is often used in conjunction with political entities, as there exist several bigrams such as *blame obama*, *vote obama*, and *hate bush*. Non-controversial comments, on the other hand, also contain terms related to politics; however, those are rather general terms such as *washington*, *politician*, and *govern* that are not specific to any political group. Note that, by definition, the set of non-controversial comments are those found at the two extremes of the spectrum defined by our Φ values. This explains why the term list for the non-controversial comments include words like *corrupt* and *hope*, which might be extracted from the comments that are either rated "negative" or "positive" by a vast majority. Another interesting example in the non-controversial term list is the word *bank*; our manual inspection of corresponding comments revealed that banks are often criticized because of their role in the financial crisis, and these comments are approved by a large majority of the users.

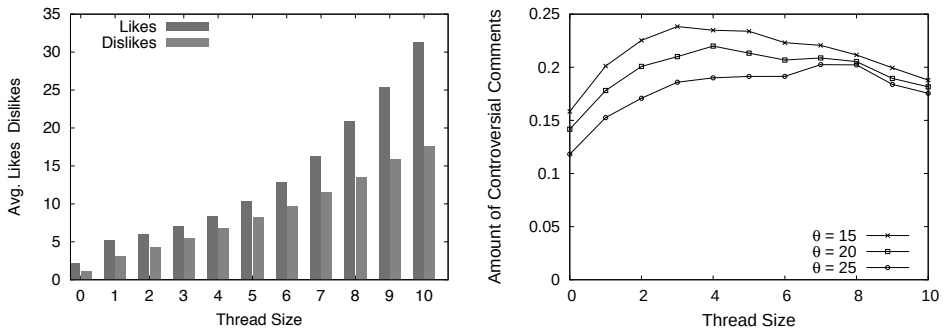
7.3. Analysis of Ratings for Comment Threads

In Section 6, we showed that comments receiving a larger number of replies have more positive ratings on the average. In this section we extend that analysis by investigating the *controversy* of comments resulting in discussion threads.

Figure 13a shows the average number of likes and dislikes of seed comments for different discussion thread sizes. We observe that, on average, seed comments for longer threads receive a larger number of likes and dislikes. Additionally, the number of likes grows faster than the number of dislikes, i.e., the comments that triggered longer discussions tend to be associated with more positive ratings. We also notice that the gap between likes and dislikes increases for larger thread sizes. Figure 13b shows the fraction of controversial seed comments (with Φ values in $[0.4, 0.6]$) with respect to the size of the initiated threads. We observe that comments initiating discussions tend to be

Table VIII: Top-20 terms according to their MI values for controversial vs. non-controversial comments.

Terms for Controversial Comments		Terms for Non-Controversial Comments	
obama	muslim	politician	need
republican	want	govern	month
liber	black	congress	law
bush	america	polit	mother
presid	right	time	help
tea	rich	money	food
parti	blame	bank	limit
gop	hate	washington	famili
2012	fact	hope	buy
democrat	believ	corrupt	day



(a) Average Likes and Dislikes (b) Fraction of Controversial Comments

Fig. 13: Comment ratings and controversy with respect to thread size.

more controversial. The fraction of controversial comments increases with thread size, but quickly saturates for threads of size 4 and above.

7.4. Temporal Characteristics of Comment Ratings

What is the average "lifetime" of a comment in terms of the community feedback it attracts? Do ratings in the earlier lifetime of a comment affect subsequent ratings? Are there preferential attachment effects, i.e. do positive/negative ratings at an early stage lead to a bias towards even more positive/negative ratings?

In order to study the temporal dynamics of comment rating and reply behavior, on the 15th of February 2013 we gathered the stories from Yahoo! News published on that day (amounting to a total of 187 news stories). For these stories, we crawled all of the available comments for a 7 day time interval, updating content of new comments and information on the temporal development of comment ratings iteratively in a round-robin fashion over the news stories. We chose a 7 day interval for the crawl as we noticed that most of the stories were removed from the system after one week. This resulted in a set of 18,902 comments being updated up to 59 times within the one week period using our data gathering strategy.

Figure 14 sketches the crawling strategy for an individual comment starting with the posting time of the comment. For each comment we defined a set of fixed time

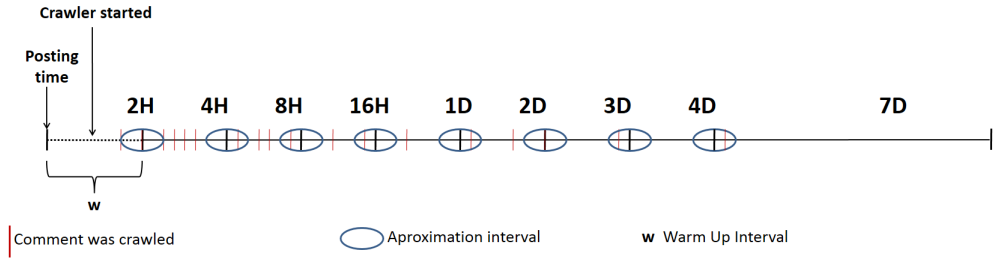


Fig. 14: Crawling strategy for temporal analysis.

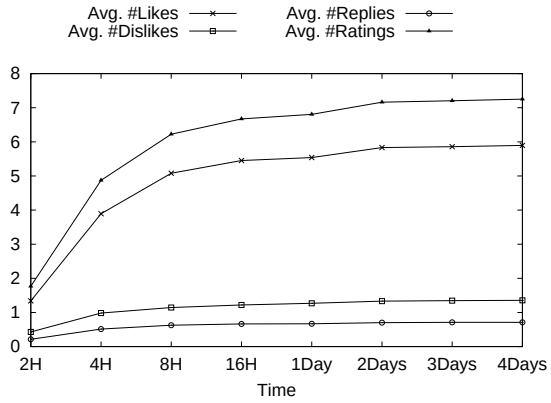


Fig. 15: Temporal evolution of average number of comment ratings, likes, and dislikes; and average number of replies.

points for updating comment rating information, corresponding to time periods of 2h (the warm-up interval), 4h, 8h, 16h, 1 day, 2 days, 3 days, and 4 days after the posting time of the comment. The red vertical lines depict the actual time when comment information was crawled and updated. Updates within a range of 20% of the fixed time periods (indicated by blue circles) where assigned to the corresponding fixed time point. This estimate was necessary due to the lack of timestamps for updates and the availability of just very rough approximations for posting times. Note that due to the difficulties implied in crawling information on posted and updated comments and limitations in the possible number of http requests per hour, time latencies occur which can result in missing some of the comments in their early life (warm-up interval) but also for some other points in time. Therefore, our final analysis was conducted on a subsample of comments fulfilling the following two criteria: 1) they were crawled in their early life (we experimentally chose a warm up interval of 2 hours), and, 2) the crawler gathered the comment rating information for the fixed time points as defined above. The final dataset used for our analysis consisted of 2,404 comments for 62 news stories.

Figure 15 shows the temporal development of the average number of likes, dislikes, ratings, and replies for comments. We observe that majority of the user interactions occur within the first 8 hours after a comment is posted. About 1 day after the posting there are clear saturation effects, indicating that comments do not receive much feedback from the community after that time period.

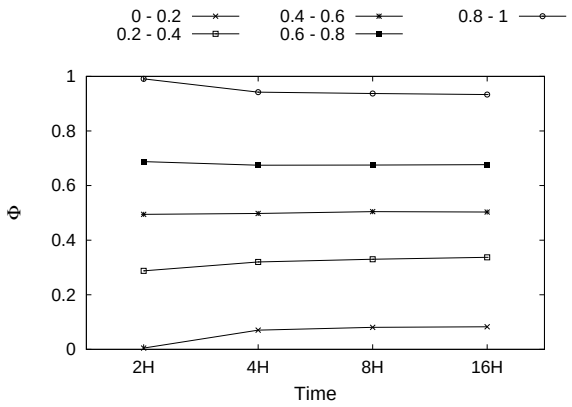


Fig. 16: Temporal evolution of average acceptance ratios for different ranges of values for initial comment acceptance ratios Φ.

In order to examine the influence of early ratings, we split the dataset of comments obtained through the time-aware crawl into five disjoint subsets depending on their ratio Φ of likes and overall ratings in the warm-up phase (i.e. the first 2h after their posting time). To this end, we split the range of possible values of Φ ([0,1]) into 5 equidistant subintervals and assigned the comments that had obtained at least one rating in the warm-up phase to the corresponding subset.

Figure 16 shows the temporal development of the average Φ values for different initial ranges of Φ. We observe that for the different starting values, the ratio Φ of likes and overall ratings stays almost constant, indicating that this ratio remains stable and that there are no preferential attachment effects.

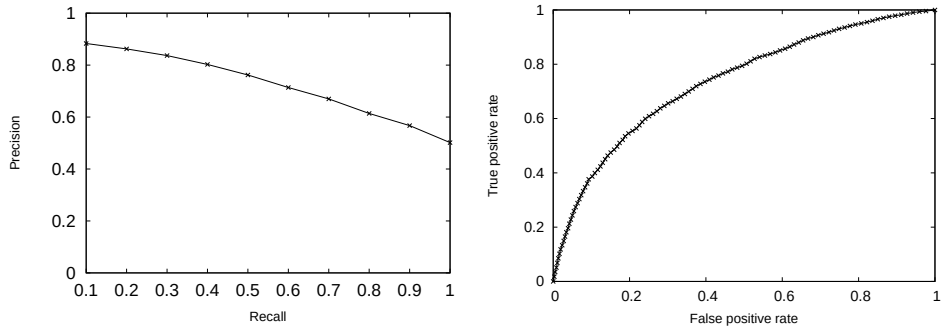
7.5. Predicting Controversial Comments

Can we predict controversial comments, i.e. comments that receive a comparable number of likes and dislikes from the community at the same time? In order to explore this, we built binary SVM classifiers based on the textual content of the comments. More specifically, from each story, we first extracted an equal number of n controversial and non-controversial comments, with n being the smaller of cardinalities of the controversial/non-controversial comments, and controversy determined by the definitions provided in Section 7.1. The overall set of controversial and non-controversial comments formed the positive and negative instances for the learning algorithm. We fixed the controversy interval in our experiments to be [0.4,0.6] (corresponding to $\delta_C = 0.1$), and varied δ_{NC} and θ to explore the impact of a wider range of Φ values for the negative class and the total number of ratings, respectively. We constructed feature vectors for comments as described in Section 5. We tested the classifier performance using 5-fold cross-validation with dataset sizes shown in Table IX (keeping 80% of the data for training and 20% for testing in each of the 5 runs) and using LIBSVM with cost parameter C=0.1. We repeated this experiment for different values of the parameters δ_{NC} and θ.

The results for the precision-recall break-even-points (BEPs) are shown in Table IX along with the total number of stories and comments used for each configuration. We can observe two main trends: First, for a given θ, increasing δ_{NC} improves the classification performance. This is reasonable, as larger δ_{NC} values restrict non-controversial comments to a narrower band where the ratio of likes and dislikes differs substantially; as a consequence, the classifier can distinguish these comments from the con-

Table IX: BEPs for controversial comment prediction.

θ	δ_{NC}=0.2			δ_{NC}=0.3			δ_{NC}=0.4		
	BEP	#stories	#comments	BEP	#stories	#comments	BEP	#stories	#comments
15	0.571	3,690	13,6000	0.59	3,439	101,000	0.649	2,861	52,000
20	0.579	2,514	57,000	0.623	2,272	42,000	0.668	1,891	24,000
25	0.589	1,752	28,000	0.633	1,564	21,000	0.679	1,258	12,000



(a) Precision-recall curve (b) ROC Curve

Fig. 17: Precision-recall and ROC curve for the classification of controversial comments (δ_{NC}=0.4).

troversial ones more easily. Secondly, as expected, comments with a higher number of ratings (corresponding to larger values of threshold θ) provide stronger evidence while learning to classify controversy.

The ROC curve is shown in Figure 17b. We obtain a value of 0.733 for the AUC (Area Under the ROC Curve) and an accuracy of 0.649. Figure 17a shows the detailed precision-recall curves for $\theta = 25$ and varying values of δ_{NC}. While the BEP values are relatively low, trading recall for precision leads to applicable results. For instance, for the non-controversial set corresponding to comments with Φ values in the first and last 10% bands (i.e., $\delta_{NC} = 0.4$), the precision is 0.859 for a recall level of 0.1 and greater than 0.8 up to a recall level of 0.3. This is useful for application scenarios such as displaying a relatively small number of potentially controversial comments at visible ranks.

7.6. Summary and Lessons Learned

In this section, we detailed our analysis of comment ratings and studied controversial comments that attract a comparable number of "likes" and "dislikes" from the community. We observed that up to 20% of the Yahoo! News comments give rise to controversies amongst raters (cf. Figure 11). Our term analysis revealed that comments centered around political entities are the most likely to create controversy. We also found that intra-thread controversy tends to increase with larger thread sizes, although it quickly saturates for threads of 5 or more comments (cf. Figure 13). Finally, while automatically distinguishing the classes of controversial and non-controversial comments turned out to be a non-trivial task, we obtained applicable results for practical scenarios by trading recall for precision (cf. Figure 17). Spotting comments likely to be controversial can be used to create ranking strategies that promote their visibility in order to increase community participation. Furthermore, controversial comment

detection could be leveraged for enhancing search user interfaces by providing new comment discovery tools.

8. COMMENT RATINGS AND POLARIZING YOUTUBE CONTENT

So far we have mainly focused on the comments themselves and have not considered the content they are associated to. While the previous section focused on *comments* that split the community, in this section we analyze polarizing *content* and the patterns it creates in the comment ratings. By "polarizing content" we mean content likely to trigger diverse opinions and sentiment, examples being content related to the war in Iraq or the presidential election in contrast to rather "neutral" topics such as chemistry or physics. Intuitively, we expect a correspondence between diverging and intensive comment rating behavior and polarizing content in YouTube. We restricted our study to YouTube as our analysis relies on tag annotations for content.

8.1. Variance of Comment Ratings as Indicator for Polarizing Videos

In order to identify polarizing videos, we computed the variance of comment ratings for each video in our dataset. Figure 18 shows examples of videos with high versus low rating variance (in our specific examples videos about an Iraqi girl stoned to death, Obama, and protest on Tiananmen Square in contrast to videos about The Beatles, cartoons, and amateur music). In order to show the relation between comment ratings and polarizing videos more systematically , we conducted a user evaluation of the top- and bottom-50 videos sorted by their variance. These 100 videos were put into random order, and evaluated by 5 users[13] on a 3-point Likert scale (3: polarizing, 1: rather neutral, 2: in between). The assessments of the different users were averaged for each video, and we computed the inter-rater agreement using the κ-measure [Rosenberg and Binkowski 2004], a statistical measure of agreement between individuals for qualitative ratings. The mean user rating for videos on top of the list was 2.085 in contrast to a mean of 1.25 for videos on the bottom (inter-rater agreement $\kappa = 0.42$); this is quite a high difference on a scale from 1 to 3, and supports our hypothesis that polarizing videos tend to trigger more diverse comment rating behavior. A t-test confirmed the statistical significance of this result (t= 7.35, d.f. = 63, $P < 0.000001$).

8.2. Variance of Comment Ratings as Indicator for Polarizing Topics

We further studied the connection between comment ratings and video tags corresponding to polarizing topics. To this end we selected all tags from our dataset occurring in at least 50 videos resulting in $1,413$ tags. For each tag we then computed the average variance of comment ratings over all videos labeled with this tag. Table X shows the top- and bottom-25 tags according to the average variance. We can clearly observe a higher tendency for tags of videos with higher variance to be associated with more polarizing topics such as *presidential*, *islam*, *irak*, or *hamas*, whereas tags of videos with low variance correspond to rather neutral topics such as *butter*, *daylight* or *snowboard*. There are also less obvious cases an example being the tag *xbox* with high rating variance which might be due to polarizing gaming communities strongly favoring either Xbox or other consoles such as PS3, another example being *f-18* with low rating variance, a fighter jet that might be discussed under rather technical aspects in YouTube (rather than in the context of wars). We evaluated this tendency in a user experiment with 3 assessors[14]. We followed the same strategy as previously described, using a 3-point Likert scale and presenting the tags to the assessors in random order. The mean user rating for tags in the top-100 of the list was 1.53 in contrast

[13]Participants consisted of PhD students and PostDocs from the institution of the first author.
[14]Again, participants consisted of PhD students and PostDocs from the affiliation of the first author.

Fig. 18: Videos with high (upper row) versus low variance (lower row) of comment ratings.

Table X: Top and Bottom-25 tags according to the variance of comment ratings for the corresponding videos.

High comment rating variance				
presidential	nomination	muslim	shakira	islam
campaign	station	itunes	grassroots	nice
xbox	barack	efron	zac	iraq
3g	kiss	obama	deals	celebrities
jew	space	shark	hamas	kiedis
Low comment rating variance				
betting	turns	puckett	tmx	tropical
skybus	peanut	defender	f-18	vlog
butter	chanukah	form	savings	iditarod
lent	daylight	egan	snowboard	havanese
menorah	casserole	1040a	1040ez	booklet

to a mean of 1.16 for tags on the bottom-100 (with inter-rater agreement $\kappa = 0.431$), supporting our hypothesis that tags corresponding to polarizing topics tend to be connected to more diverse comment rating behavior. The statistical significance of this result was confirmed by a t-test (t=4.86, d.f. = 132, P = 0.0000016).

In order to study topics beyond individual tags and to obtain more context-related information, we additionally employed Latent Dirichlet Allocation (LDA) [Blei et al. 2003] and modeled each tag-based representation of a video as a mixture of latent topics[15]. In a nutshell, given a set of term sets (videos v_i represented by their tags in our case) and the desired number of latent topics, k, LDA outputs the probabilities $P(z_j|v_i)$ that topic z_j is contained in video v_i. In addition, LDA computes term probabilities $P(t_j|z_i)$ for tags t_j; the terms with the highest probabilities for a latent topic z_i can be used to represent that topic.

We empirically chose the number of latent topics as 200 for our YouTube dataset. Analogously to our method for identifying terms related to polarizing topics, we com-

[15]We used the LDA implementation in the Mallet library at http://mallet.cs.umass.edu/.

Table XI: Most probable terms for the top-5 and bottom-5 latent topics according to the comment rating variance of the corresponding videos. Topics for videos with high comment rating variance include *politics* and *terrorism*; for low comment rating variance topics found include *holidays* and *taxes*.

Top-5 topics connected to videos with high comment rating variance				
TOPIC 1:	TOPIC 2:	TOPIC 3:	TOPIC 4:	TOPIC 5:
saddam	vegas	shakira	de	clinton
hussein	simpson	lie	mayo	obama
iraq	las	hips	cinco	barack
pacing	oj	don	san	bill
dr	peanut	dont	mexico	hillary
pluto	butter	wolf	mexican	president
hanging	recall	hurley	diego	john
stc	prison	elizabeth	jose	election
division	trial	dance	california	bush
healthcare	talent	live	latino	politics
Bottom-5 topics connected to videos with low comment rating variance				
TOPIC 1:	TOPIC 2:	TOPIC 3:	TOPIC 4:	TOPIC 5:
kurt	tax	easter	dance	carey
vonnegut	taxes	chanukah	girl	mariah
language	income	bunny	hot	mary
arts	irs	egg	blonde	ron
book	aid	menorah	katie	lol
science	form	gas	babe	porn
theatre	free	jewish	big	funny
learn	forms	prices	ass	cannon
humanities	video	eggs	cindy	fail
art	federal	darfur	black	awesome

puted the average variance of comment ratings over all videos that belong to a latent topic, weighting the contributions of the videos by their probability values $P(z_j|v_i)$. Table XI shows the top-5 and bottom-5 latent topics (represented by their most probable terms) according to their average variance scores. Similar as for individual terms, polarizing and neutral topics can be successfully distinguished. In particular, the topics that are centered around *Iraq*, *O.J. Simpson trial* and *American politics* (i.e, the first, second and last columns of top-5 topics in Table XI, respectively) are obviously polarizing. On the other hand, the bottom-5 topics look rather neutral, being related to the issues like *Kurt Vonnegut* (an American writer), *tax forms*, *girls*, etc.

8.3. Summary and Lessons Learned

This section focused on the relationship between the polarizing characteristics of content and the commenting behavior on YouTube. We conducted a user study where we found evidence that high variance in comment ratings is correlated to a high video polarity. Furthermore, we grouped videos by tag and repeated the experiment, finding that tags for videos with high rating variance tend to be associated with more polarizing topics. We made similar observations for latent topics which we analyzed in order to capture a broader context that goes beyond individual terms. Videos with polarizing content have the potential of engaging users in interesting discussions, increasing participation and, thus, enriching the community.

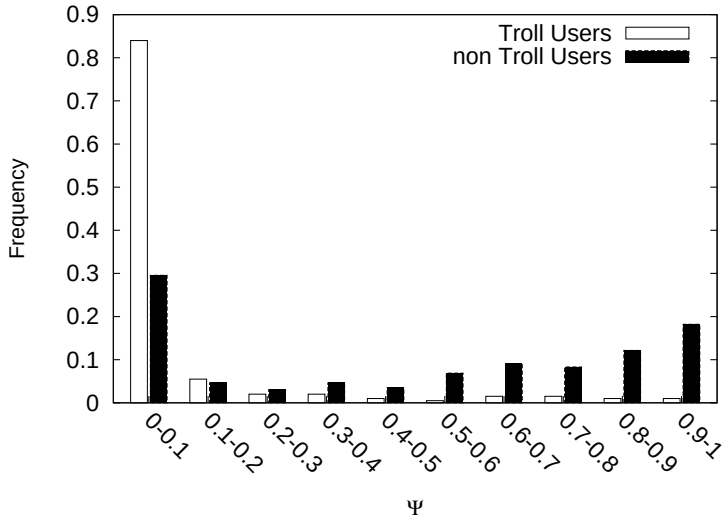


Fig. 19: Distribution of troll and non-troll users in YouTube with respect to user approval ratio (Ψ) intervals.

9. ANALYSIS AND DETECTION OF TROLL USERS

In this section, we will focus on the *users* commenting on content in social web environments. Commenting tools in social websites are mostly used to share legitimate opinions and feelings. Nevertheless, it is also common to find users that abuse this mechanism in various ways. These include posting links to external web pages aiming at increasing their visibility (i.e. spamming), or "posting disruptive, false or offensive comments to fool and provoke other users" (i.e. trolling [J.Kunegis and C.Bauckhage 2009]). We conduct an exploratory analysis of the presence of troll users (*trolls*) in social websites, and study methods for automatically detecting potential trolls based on the textual content of their comment history.

9.1. Finding Trolls

Yahoo! News allows participants to use non-unique identifiers for commenting, which renders the task of modeling troll behavior unfeasible. Hence, we used the YouTube collection as our main data source for our analysis. Given this limitation of Yahoo! News, we decided to collect an additional dataset in order to provide a more comprehensive analysis of trolling characteristics in social websites. In particular, we crawled the popular technology news website Slashdot[16], as for this site it is possible to easily obtain a set of manually classified trolls and their comments. To this end, we followed the procedure proposed by [J.Kunegis and C.Bauckhage 2009] and extracted a set of troll users from a special user account, called *No More Trolls*, which tags all *known* trolls as its *foes* to help other users avoid them. We crawled the 24 most recent comments (i.e. the maximum number of comments per user shown by Slashdot) from all users listed as trolls. The resulting collection includes 200 users and 4310 comments. An additional random sample of 200 users not contained in the *No More Trolls* list was crawled to represent the negative class, i.e. "non-trolls".

[16]http://slashdot.org/

Extracting a comparable number of trolls from the YouTube dataset is not straightforward. First, troll detection requires manually assessing the content of each user's comments as YouTube does not provide a list of troll users flagged by the community. Second, the proportion of trolls is significantly lower than that of legitimate users [J.Kunegis and C.Bauckhage 2009]. Therefore, identifying a comparable amount of trolls in YouTube using a random sampling strategy would require manually inspecting comments from thousands of users. To decrease the manual effort required, we used a simple heuristic to increase the chance of finding trolls in our sample by means of the *user approval ratio* $\Psi := \frac{pos(u)}{pos(u)+neg(u)}$, where $pos(u)$ and $neg(u)$ denote the number of positively and negatively rated comments for a given user u, respectively. Low values of this ratio indicate strong rejection by the community for the comments of a particular user, while high values indicate general acceptance of the user's opinions. We used this metric to sample YouTube users by randomly selecting 500 users with $\Psi(u) \in [0, 0.1]$ under the assumption that a significant number of trolls would fall into this interval. In order to obtain a set containing more non-troll users we also sampled a set of 500 users with approval ratio $\Psi(u) \in [0.1, 1]$. The final set of 1,000 users was then *manually* annotated by one of the authors using the following three labels based on the content of their comments: "troll", "non-troll", or "unknown".

9.2. Trolls and Community Ratings

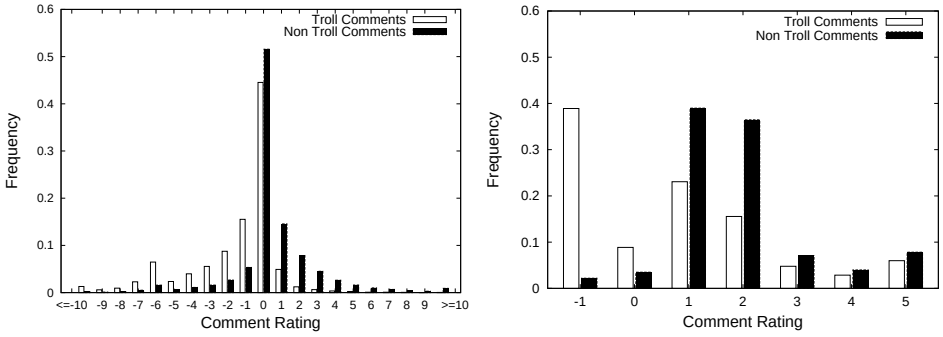
Can comment ratings serve as an indicator for trolling behavior? Figure 19 shows the distribution of troll and non-troll users in YouTube with respect to the user approval ratios Ψ. This figure clearly illustrates the higher proportion of trolls found in the $[0, 0.1]$ Ψ range, as compared with the proportion at higher levels of Ψ. This result provides empirical support for the heuristic chosen in our sampling strategy. We also observe a large percentage of trolls in the $[0.1, 0.2]$ range, whereas just a tiny fraction of users are found to be trolls for $\Psi > 0.2$. This confirms the intuition that comment ratings serve as good proxies for troll identification in online communities.

Figure 20 shows the distribution of comment ratings from YouTube (Figure 20a) and Slashdot (Figure 20b). Note that our sampling strategy for detecting trolls in YouTube is biased towards low rated comments, as 50% of the comments were chosen from Ψ values in the range $[0, 0.1]$. As illustrated in Figure 19, this bias significantly affects the distribution of non-troll comment ratings, but has just a marginal effect on the distribution of troll comment ratings as they mostly feature low rating values. Therefore, we address our sampling bias by comparing the ratings of comments from trolls in this sample with ratings from comments in the whole dataset (including troll and non-troll users). Both plots show a clear trend of comments from troll users having lower ratings than comments from non-troll users in both communities.

9.3. Content-based Troll Prediction

How does vocabulary usage differ for troll and non-troll users, and can the textual content of comments be leveraged for detecting trolls?

We compared the most discriminative terms of the comments from troll and non-troll users in YouTube and Slashdot. For each dataset, we randomly selected 200 troll and 200 non-troll users and extracted 24 comments sampled uniformly at random. Analogously to the term analyses in previous sections we computed the top-ranked (stemmed) terms with respect to the Mutual Information measure. Table XII reveals that the terms used in comments from trolls are very similar for YouTube and Slashdot, and these terms are mostly offensive. Despite exhibiting less similarity, the term lists for non-troll users generally include more positive terms such as *good*, *love*, *beauty* (in YouTube) and *like*, *pretti*, *agre* (in Slashdot). Some of the Slashdot terms look coun-



(a) YouTube (b) Slashdot

Fig. 20: Comment rating distribution for comments from troll users and non-troll users in (a) YouTube and, (b) Slashdot.

Table XII: Top-20 terms according to their MI values for troll vs. non-troll comments.

Terms for Troll Comments				**Terms for Non-Troll Comments**			
YouTube		Slashdot		YouTube		Slashdot	
fuck	dick	fuck	bush	love	happen	use	pretti
shit	stupid	post	vomit	look	use	think	peopl
suck	young	troll	failur	good	thought	work	time
ass	hey	slashdot	nigger	miss	great	http	agre
white	cunt	linux	enjoy	time	doe	year	game
nigger	black	shit	ass	awesom	thank	thing	look
bitch	retard	fail	love	think	clay	problem	actual
free	cock	die	cybernet	agre	hot	compani	phone
gay	watch	gay	crapflood	lol	end	doe	realli
u	jew	fp	clit	s	govern	know	probabl

terintuitive at first sight. For instance, terms used by trolls in Slashdot include *linux*, *slashdot*, or *cybernet*. Further inspection of the data revealed that trolls often use them as their target for complains and insults (e.g. *"linux is a failure"*). On the other hand, we notice that, other than in YouTube, terms in the non-troll category seem to be slightly less positive (or even neutral) in Slashdot (e.g. *http*, *game*, *phone*). This could be related to the fact that comments in Slashdot are mostly used to engage in technical discussions. An interesting observation is that *http* is a discriminative term for non-troll comments in Slashdot; this mainly corresponds to posting links in comments which are often appreciated by the community.

The difference in the terms chosen by trolls and normal users within their comments encouraged us to study the possibility of training SVM classifiers (using LIBSVM with linear kernel and cost parameter C=0.1) to predict trolls by using the textual content of their comments. Comments from one user were merged into a single "virtual" comment; feature vectors were then constructed as described in Section 5. We used 200 users (and 4,800 comments) for each class of users, and a 50-50 split with 2-fold cross validation[17] to report the average classification performance.

[17]We used 2-fold CV due the the relatively small size of the datasets. Splitting the set into a larger number of folders would have rendered the computation of the individual precision-recall curves in each CV run very

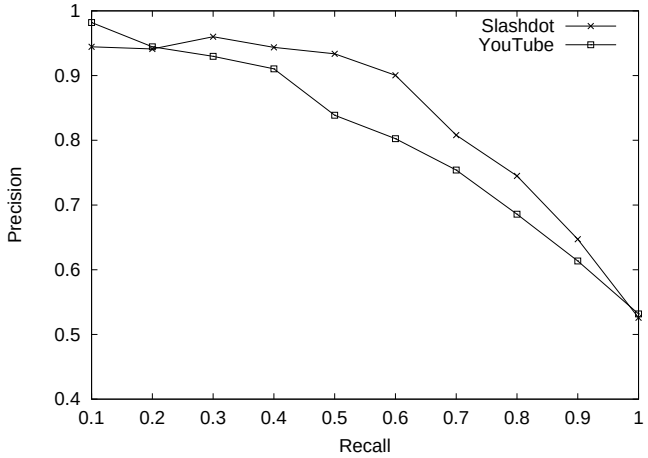


Fig. 21: Precision-recall curves for troll detection in the YouTube and Slashdot datasets.

Figure 21 shows the precision-recall curves for predicting troll users from YouTube and Slashdot. We observe BEP values of 0.682 and 0.742 for YouTube and Slashdot, respectively. Our findings reveal that the precision is greater than 0.8 up to a recall value of 50% for both datasets. The relatively large difference in the classification performance for the datasets suggests inherent differences in the communities and their commenting behavior, as previously observed in Section 3. The YouTube collection mainly contains short and unelaborate opinions that provide fewer cues for the correct classification of users as trolls.

Note that troll detection applications should be tuned to seek high precision. Automatic troll detection needs to avoid censoring legitimate users, as this could result in user frustration and, ultimately, community destruction. We believe that troll detection should be carefully assessed by human supervisors to avoid any possibility of user loss, as users are the main asset of social websites.

9.4. Summary and Lessons Learned

In this section we analyzed troll users in the context of comments. In addition to analyzing YouTube, we also examined trolls in the technical discussion forum Slashdot. We discovered that for both platforms the community ratings for comments of troll users clearly differ from those of other users (cf. Figure 19 and 20). More specifically , the approval ratio of comment ratings for troll users is substantially lower. This indicates that comment ratings can be a useful source of information for detecting trolls. Differences between comments from troll and non-troll users and similarities with terms and sentiments described in earlier sections are further reflected by the results of our discriminative term analysis (cf. Table XII). The observed patterns motivated the application of machine learning for troll detection based on the content of user comments. Results of the classification experiments (cf. Figure 21) are promising and demonstrate the potential for automatically identifying likely candidates for trolls in social platforms.

unstable (especially due to the very small number of documents that would be involved for computing the low-recall part of the curves).

10. CONCLUSIONS AND DISCUSSION

In this paper we provided an in-depth analysis of user comments in two prominent social websites, YouTube and Yahoo! News, aiming at achieving a better understanding of community feedback on the Social Web. For troll detection we conducted additional experiments on Slashdot as this dataset has been explored in this context in previous studies. There are interesting aspects to the contrast of different online communities, as this enables us to contextualize results, understand what makes each community unique and also what elements they share. This latter aspect is especially relevant, given the differences between the media supported by these datasets (videos vs. news stories), the nature of the content (user generated vs. editorially prepared), the different topics covered, and the nature of the user communities attracted.

The studies presented in this paper focused on understanding the users and content in online communities, and discussing implications for designing collaborative systems that promote and encourage participation and enhance overall user experience. In this section we summarize our major findings and relate them to the high-level research goals that guide this paper. We also discuss how these findings can foster further research and be leveraged to improve specific applications.

10.1. Understanding community feedback and meta-feedback

Two main results derive from the presented studies:

— Community meta-feedback provided through comment ratings is indeed dependent on characteristics of the comments content such as orientation of opinions; we observed that positive opinions expressed in the comments attract positive community feedback, and vice versa.
— Comment content helps predicting various types of community feedback, such as overall comment rating, ratio of likes and dislikes for the comments, likelihood of comments triggering replies, and further participation from the community.

These two results are clearly more prominent in the YouTube community, where the abuse of language occurs significantly more often as compared to Yahoo! News, partly because of stricter comment filtering policies in the Yahoo! News system.

These findings provide important design guidelines for building more engaging and usable online communities. Consider the vast amount of comments attracted by popular content; our results show that we can infer the potential acceptance of comments, making it possible to enhance relevant comment discovery by highlighting the ones that are likely to be the most liked, disliked, or both (in the case of controversial comments) as well as the most replied, which are likely to trigger interesting discussions. Furthermore, these results could be used to provide multiple facets (e.g. predicted controversy or community acceptance) to complement pure textual queries. In addition, comment search engines could leverage our results for deriving ranking schemes that promote comments based on their ability to attract replies and votes, as a way to give visibility to a wider range of opinions.

Smart methods for automatic comment curation and discovery have the main benefit of increasing user participation by making relevant content more visible. In addition, they may serve to enhance the community in various other ways, for instance by promoting comments that spark new ideas or enrich the platform by offering alternative views. Furthermore, the analyses proposed in this paper can serve as the basis for developing methods to better identify what opinions are supported or rejected by the community, as well as which ones trigger controversy and discussions or, in general, carry valuable information both academically (for social scientists) and practically (for marketing experts, etc.). For instance, journalists interested in rapidly understanding

the range of opinions about topics of public interest could use the proposed controversy search facet to this end. This can be also applied for understanding community dynamics and studying the evolution of controversy along time.

Our results also showed the potential impact of interface design decisions on user participation. In this regard, Yahoo! News shows a clear tendency towards reply comments not attracting ratings from the community, which we identified by contrasting its reply distribution with YouTube's. The user interface hides such replies by default, effectively requiring action from the community to unhide them before they can be read and rated. This effect is not present in YouTube, where replies to comments are not hidden by default.

10.2. Understanding the content

The main result connecting comments and content is that community feedback (and its variance) on comments can help identifying polarizing content, that is, content that generates rich discussions between community members with contrary opinions. Polarizing content has the potential of engaging users in interesting discussions, increasing participation and, thus, enriching the community.

We discuss how the variance of ratings for a specific piece of content serves as a strong social signal to pinpoint polarizing content. This approach can either replace or complement other methods for estimating the degree of polarization, either based on the analysis of:

— the actual content, which can be challenging especially in the case of multimedia documents, or
— its metadata, which in the general case is noisy and lacks sufficient detail to produce accurate results.

10.3. Understanding the user

We finally studied the characteristics of users, specifically trolls, commenting on content in social web environments. We found comment content can be leveraged to effectively identify troll users. Trolls compromise the usability and overall experience of online communities by deliberately taking actions to change their organic social dynamics. These actions can, therefore, have a negative impact in both the quality of community interaction as well as the effectiveness of further analyses to be conducted on the accumulated social data. Detecting troll users plays an important role in improving experience and increasing user engagement within the online community.

In this regard, mining comments from registered users enables the inference of automatic user profiles that can be effectively exploited for troll detection. Enabling anonymous comments could promote further community participation, but at the same time potentially harm the stability of the site as troll behavior is more difficult to recognize, track, and prevent. Both aspects need to be considered and reflected upon when designing such collaborative websites.

REFERENCES

AGICHTEIN, E., CASTILLO, C., DONATO, D., GIONIS, A., AND MISHNE, G. 2008. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*. WSDM '08. ACM, New York, NY, USA, 183–194.

ALONZO, M. AND AIKEN, M. 2004. Flaming in electronic communication. *Decis. Support Syst. 36,* 3, 205–213.

BLEI, D. M., NG, A. Y., AND JORDAN, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res. 3*, 993–1022.

CHA, M., KWAK, H., RODRIGUEZ, P., AHN, Y.-Y., AND MOON, S. 2007. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. IMC '07. ACM, New York, NY, USA, 1–14.

CHANG, C. AND LIN, C. 2011. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2011 2*, 27:1–27:27.

CHELARU, S., ORELLANA-RODRIGUEZ, C., AND ALTINGOVDE, I. S. 2012. Can social features help learning to rank youtube videos? In *Proceedings of the 13th International Conference on Web Information Systems Engineering*. WISE '12. 552–566.

CHENG, X., DALE, C., AND LIU, J. 2007. Understanding the characteristics of internet short video sharing: Youtube as a case study. In *Technical Report arXiv:0707.3670v1 cs.NI*. Cornell University, arXiv e-prints, New York, NY, USA.

DALAL, O., SENGEMEDU, S. H., AND SANYAL, S. 2012. Multi-objective ranking of comments on web. In *Proceedings of the 21st international conference on World Wide Web*. WWW '12. ACM, New York, NY, USA, 419–428.

DANESCU-NICULESCU-MIZIL, C., KOSSINETS, G., KLEINBERG, J., AND LEE, L. 2009. How opinions are received by online communities: a case study on amazon.com helpfulness votes. In *WWW '09: Proceedings of the 18th international conference on World wide web*. ACM, New York, NY, USA, 141–150.

DE CHOUDHURY, M., SUNDARAM, H., JOHN, A., AND SELIGMANN, D. D. 2009. What makes conversations interesting?: themes, participants and consequences of conversations in online social media. In *Proceedings of the 18th international conference on World wide web*. WWW '09. ACM, New York, NY, USA, 331–340.

DENECKE, K. 2008. Using sentiwordnet for multilingual sentiment analysis. In *Proceedings of the 24th International Conference on Data Engineering Workshops*. 507–512.

DUMAIS, S., PLATT, J., HECKERMAN, D., AND SAHAMI, M. 1998. Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management*. CIKM '98. ACM, New York, NY, USA, 148–155.

ESULI, A. 2008. Automatic generation of lexical resources for opinion mining: models, algorithms and applications. *SIGIR Forum 42*, 105–106.

ESULI, A. AND SEBASTIANI, F. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation LREC '06*. 417–422.

FELLBAUM, C., Ed. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

FILIPPOVA, K. AND HALL, K. B. 2011. Improved video categorization from text metadata and user comments. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. SIGIR '11. ACM, New York, NY, USA, 835–842.

GILL, P., ARLITT, M., LI, Z., AND MAHANTI, A. 2007. Youtube traffic characterization: a view from the edge. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. IMC '07. ACM, New York, NY, USA, 15–28.

GÓMEZ, V., KALTENBRUNNER, A., AND LÓPEZ, V. 2008. Statistical analysis of the social network and discussion threads in slashdot. In *Proceedings of the 17th international conference on World Wide Web*. WWW '08. ACM, New York, NY, USA, 645–654.

GÓMEZ, V., KAPPEN, H., LITVAK, N., AND KALTENBRUNNER, A. 2012. A likelihood-based framework for the analysis of discussion threads. *World Wide Web*, 1–31.

HANNA, R., ROHM, A., AND CRITTENDEN, V. L. 2011. We're all connected: The power of the social media ecosystem. *Business Horizons 54,* 3, 265–273.

HARPER, F. M., RABAN, D., RAFAELI, S., AND KONSTAN, J. A. 2008. Predictors of answer quality in online q&a sites. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*. CHI '08. ACM, New York, NY, USA, 865–874.

HARWOOD, A. M. AND HAHN, C. L. 1990. *Controversial Issues in the Classroom*. ERIC Clearinghouse for Social Studies/Social Science Education.

HSU, C., KHABIRI, E., AND CAVERLEE, J. 2009. Ranking comments on the social web. In *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04*. 90–97.

HU, M., SUN, A., AND LIM, E.-P. 2008. Comments-oriented document summarization: understanding documents with readers' feedback. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 291–298.

J.KUNEGIS, A. AND C.BAUCKHAGE. 2009. The slashdot zoo: Mining a social network with negative edges. *ACM Transactions on Intelligent Systems and Technology, 2011*.

JOACHIMS, T. 1998. Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*. ECML '98. Springer-Verlag, London, UK, UK, 137–142.

KIETZMANN, J. H., HERMKENS, K., MCCARTHY, I. P., AND SILVESTRE, B. S. 2011. Social media? get serious! understanding the functional building blocks of social media. *Business Horizons 54,* 3, 241 – 251.

KIM, S.-M., PANTEL, P., CHKLOVSKI, T., AND PENNACCHIOTTI, M. 2006. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. EMNLP '06. Association for Computational Linguistics, Stroudsburg, PA, USA, 423–430.

KUYPERS, J. A. 2002. *Press Bias and Politics: How the Media Frame Controversial Issues*. Praeger.

LI, Q., WANG, J., CHEN, Y. P., AND LIN, Z. 2010. User comments for news recommendation in forum-based social media. *Information Sciences 180,* 24, 4929–4939.

LU, Y., ZHAI, C., AND SUNDARESAN, N. 2009. Rated aspect summarization of short comments. In *Proceedings of the 18th international conference on World wide web*. WWW '09. ACM, New York, NY, USA, 131–140.

MANNING, C. AND SCHUETZE, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

MISHNE, G. AND GLANCE, N. 2006. Leave a reply: An analysis of weblog comments. In *Third annual workshop on the Weblogging ecosystem*. Edinburgh, Scotland.

MISHRA, A. AND RASTOGI, R. 2012. Semi-supervised correction of biased comment ratings. In *Proceedings of the 21st international conference on World Wide Web*. WWW '12. ACM, New York, NY, USA, 181–190.

PANG, B., LEE, L., AND VAITHYANATHAN, S. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*. EMNLP '02. Association for Computational Linguistics, Stroudsburg, PA, USA, 79–86.

PARK, S., KO, M., KIM, J., LIU, Y., AND SONG, J. 2011. The politics of comments: predicting political orientation of news stories with commenters' sentiment patterns. In *Proceedings of the ACM 2011 conference on Computer Supported Cooperative Work*. 113–122.

POTTHAST, M., STEIN, B., LOOSE, F., AND BECKER, S. 2012. Information retrieval in the commentsphere. *ACM Trans. Intell. Syst. Technol. 3,* 4, 68:1–68:21.

ROSENBERG, A. AND BINKOWSKI, E. 2004. Augmenting the kappa statistic to determine interannotator reliability for multiply labeled data points. In *Proceedings of HLT-NAACL 2004: Short Papers*. HLT-NAACL-Short '04. Association for Computational Linguistics, Stroudsburg, PA, USA, 77–80.

ROWE, M., ANGELETOU, S., AND ALANI, H. 2011a. Anticipating discussion activity on community forums. In *Proceedings of the PASSAT/SocialCom*. 315–322.

ROWE, M., ANGELETOU, S., AND ALANI, H. 2011b. Predicting discussions on the social semantic web. In *Proceedings of the 8th extended semantic web conference on The semanic web: research and applications - Volume Part II*. ESWC'11. Springer-Verlag, Berlin, Heidelberg, 405–420.

SAN PEDRO, J., YEH, T., AND OLIVER, N. 2012. Leveraging user comments for aesthetic aware image search reranking. In *Proceedings of the 21st international conference on World Wide Web*. WWW '12. ACM, New York, NY, USA, 439–448.

SCHUTH, A., MARX, M., AND DE RIJKE, M. 2007. Extracting the discussion structure in comments on news-articles. In *Proceedings of the 9th annual ACM international workshop on Web information and data management*. WIDM '07. ACM, New York, NY, USA, 97–104.

SHMUELI, E., KAGIAN, A., KOREN, Y., AND LEMPEL, R. 2012. Care to comment?: recommendations for commenting on news stories. In *Proceedings of the 21st international conference on World Wide Web*. WWW '12. ACM, New York, NY, USA, 429–438.

SIERSDORFER, S., CHELARU, S., NEJDL, W., AND SAN PEDRO, J. 2010. How useful are your comments?: analyzing and predicting youtube comments and comment ratings. In *Proceedings of the 19th international conference on World wide web*. WWW '10. ACM, New York, NY, USA, 891–900.

SIERSDORFER, S., SAN PEDRO, J., AND SANDERSON, M. 2009. Automatic video tagging using content redundancy. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, New York, NY, USA, 395–402.

SUSARLA, A., OH, J.-H., AND TAN, Y. 2012. Social networks and the diffusion of user-generated content: Evidence from youtube. *Info. Sys. Research 23,* 1, 23–41.

TATAR, A., LEGUAY, J., ANTONIADIS, P., LIMBOURG, A., DE AMORIM, M. D., AND FDIDA, S. 2011. Predicting the popularity of online articles based on user comments. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*.

THELWALL, M., SUD, P., AND VIS, F. 2012. Commenting on youtube videos: From guatemalan rock to el big bang. *J. Am. Soc. Inf. Sci. Technol. 63,* 3, 616–629.

THOMAS, M., PANG, B., AND LEE, L. 2006. Get out the vote: determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. EMNLP '06. Association for Computational Linguistics, Stroudsburg, PA, USA, 327–335.

TSAGKIAS, M., WEERKAMP, W., AND DE RIJKE, M. 2010. News comments: Exploring, modeling, and online prediction. In *Proceedings of the 32nd European Conference on IR Research*. 191–203.

VELOSO, A., JR., W. M., MACAMBIRA, T., GUEDES, D., AND ALMEIDA, H. 2007. Automatic moderation of comments in a large on-line journalistic environment. In *Proceedings of the International Conference on Weblogs and Social Media*.

WANG, C., YE, M., AND HUBERMAN, B. A. 2012. From user comments to on-line conversations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '12. ACM, New York, NY, USA, 244–252.

WEIMER, M., GUREVYCH, I., AND MHLHUSER, M. 2007. Automatically assessing the post quality in online discussions on software. In *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. 125–128.

WU, F. AND HUBERMAN, B. A. 2008. How public opinion forms. In *Proceedings of the 4th International Workshop on Internet and Network Economics*. WINE '08. Springer-Verlag, Berlin, Heidelberg, 334–341.

YANG, Y. AND PEDERSEN, J. O. 1997. A comparative study on feature selection in text categorization. Morgan Kaufmann Publishers, 412–420.

YANO, T. AND SMITH, N. A. 2010. What's worthy of comment? content and comment volume in political blogs. In *Proceedings of the Fourth International Conference on Weblogs and Social Media*.

YEE, W. G., YATES, A., LIU, S., AND FRIEDER, O. 2009. Are web user comments useful for search? In *Proceedings of the SIGIR' 09 Workshop on LSDS-IR*.