How Useful are Your Comments? Analyzing and Predicting YouTube Comments and Comment Ratings

Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl
L3S Research Center
Appelstr. 9a
30167 Hannover, Germany
{siersdorfer, chelaru, nejdl}@L3S.de

Jose San Pedro
Telefonica Research
Via Augusta, 171
Barcelona 08021, Spain
jsanpedro@mac.com

ABSTRACT

An analysis of the social video sharing platform YouTube reveals a high amount of community feedback through comments for published videos as well as through meta ratings for these comments. In this paper, we present an in-depth study of commenting and comment rating behavior on a sample of more than 6 million comments on 67,000 YouTube videos for which we analyzed dependencies between comments, views, comment ratings and topic categories. In addition, we studied the influence of sentiment expressed in comments on the ratings for these comments using the SentiWordNet thesaurus, a lexical WordNet-based resource containing sentiment annotations. Finally, to predict community acceptance for comments not yet rated, we built different classifiers for the estimation of ratings for these comments. The results of our large-scale evaluations are promising and indicate that community feedback on already rated comments can help to filter new unrated comments or suggest particularly useful but still unrated comments.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Experimentation, Measurement

Keywords

comment ratings, community feedback, youtube

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 mechanisms for supporting collaboration, and attractiveness of shared annotated material (images in Flickr, bookmarks in del.icio.us, etc.). For video sharing, the most popular site is YouTube[1]. Recent studies have shown that traffic to/from this site accounts for over 20% of the web total and 10% of the whole internet [3], and comprises 60% of the videos watched on-line [11].

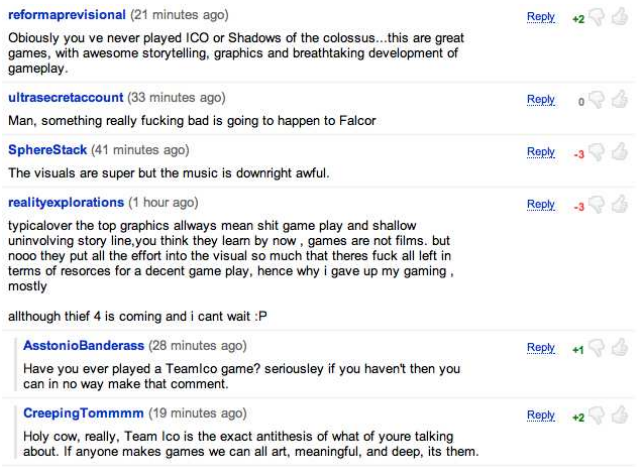

Figure 1: Comments and Comment Ratings in YouTube

YouTube provides several social tools for community interaction, including the possibility to comment published videos and, in addition, to provide ratings about these comments by other users (see Figure 1). These meta ratings serve the purpose of helping the community to filter relevant opinions more efficiently. Furthermore, because negative votes are also available, comments with offensive or inappropriate content can be easily skipped.

The analysis of comments and associated ratings constitutes a potentially interesting data source to mine for obtaining implicit knowledge about users, videos, categories and community interests. In this paper, we conduct a study of this information with several complementary goals. On the one hand, we study the viability of using comments and community feedback to train classification models for deciding on the likely community acceptance of new comments. Such models have direct application to the enhancement of comment browsing, by promoting interesting comments even in the absence of community feedback. On the other hand, we perform an in-depth analysis of the distribution of comment ratings, including qualitative and quantitative studies about

[1] http://www.youtube.com

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

sentiment values of terms and differences across categories. Can we predict the community feedback for comments? Is there a connection between sentiment and comment ratings? Can comment ratings be an indicator for polarizing content? Do comment ratings and sentiment depend on the topic of the discussed content? These are some of the questions we investigate in this paper by analyzing a large sample of comments from YouTube.

Clearly, due to the continuing and increasing stream of comments in social sharing environments such as YouTube, the community is able to read and rate just a fraction of these. The methods we present in this paper can help to automatically structure and filter comments. Analyzing the ratings of comments for videos can provide indicators for highly polarizing content; users of the system could be provided with different views on that content using comment clustering and aggregation techniques. Furthermore, automatically generated content ratings might help to identify users showing malicious behavior such as spammers and trolls at an early stage, and, in the future, might lead to methods for recommending to an individual user of the system other users with similar interests and points of views.

The rest of this paper is organized as follows: In Section 2 we discuss related work on user generated content, product reviews and comment analysis. Section 3 describes our data gathering process, as well as the characteristics of our dataset. In Section 4 we analyze the connection between sentiment in comments and community ratings using the SentiWordNet thesaurus. We then provide a short overview of classification techniques in Section 5, explain how we can apply these techniques to rate comments, and provide the results of large-scale classification experiments on our YouTube data set. In Section 6 we analyze the correspondence between comment ratings and polarizing content through user experiments. Section 7 describes dependencies of ratings and sentiments on topic categories. We conclude and show directions for future work in Section 8.

2. RELATED WORK

There is a body of work on analyzing product reviews and postings in forums. In [4] 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.* [17] use a latent topic approach to extract rated quality aspects (corresponding to concepts such as "price" or "shipping") from comments in ebay. In [27] the temporal development of product ratings and their helpfulness and dependencies on factors such number of reviews or effort required (writing 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 "priori effect" (e.g. Did the inquirer some apriori research on the topic?) is manually analyzed in [12]. In comparison, our paper focuses on *community ratings for comments and discussions* rather than product ratings.

Work on sentiment classification and opinion mining such as [19, 25] 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. Recent work in this area makes also use of SentiWordNet [5] to improve classification performance. 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 plethora of work on classification using probabilistic and discriminative models [2] and learning regression and ranking functions [24, 20, 1]. The popular SVM Light software package [14] provides various kinds of parameterizations and variations of SVM training (e.g., binary classification, SVM regression and ranking, transductive SVMs, etc.). In this paper we will apply these techniques in a novel context to automatic classification of comment acceptance.

Kim *et al.* [15] 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. Weimer *et al.* [26] make use of a similar idea to automatically predict the quality of posts in the software online forum Nabble.com. Liu *et al* [16] describe an approach for aggregation of ratings on product features using helpfulness classifiers based on a manually determined ground truth, and compare their summarization with special "editor reviews" on these sites. Another example of using community feedback to obtain training data and ground truth for classification and regression can be found in our own work [22], for an entirely different domain, where tags and visual features in combination with favorite assignments in Flickr are used to classify and rank photos according to their attractiveness.

Compared to previous work, our paper is the first to apply and evaluate automatic classification methods for comment acceptance in YouTube. Furthermore, we are the first to provide an in-depth analysis of the distribution of YouTube comment ratings, including both qualitative and quantitative studies as well as dependencies on comment sentiment, rating differences between categories, and polarizing content.

3. DATA

We created our test collection by formulating queries and subsequent searches for "related videos", analogously to the typical user interaction with the YouTube system. Given that an archive of most common queries does not exist for YouTube, we selected our set of queries from Google's Zeitgeist archive from 2001 to 2007, similarly to our previous work [23]. These are generic queries, used to search for web pages. In this way, we obtained 756 keyword queries.

In 2009, for each video we gathered the first 500 comments (if available) for the video, along with their authors, timestamps and comment ratings. YouTube computes comment ratings by counting the number of "thumbs up" or "thumbs down" ratings, which correspond to positive or a negative votes by other users. In addition, for each video we collected meta data such as title, tags, category, description, upload date as well as statistics provided by YouTube such as overall number of comments, views, and star rating for the video. The complete collection used for evaluation had a final size of $67,290$ videos and about 6.1 million comments.

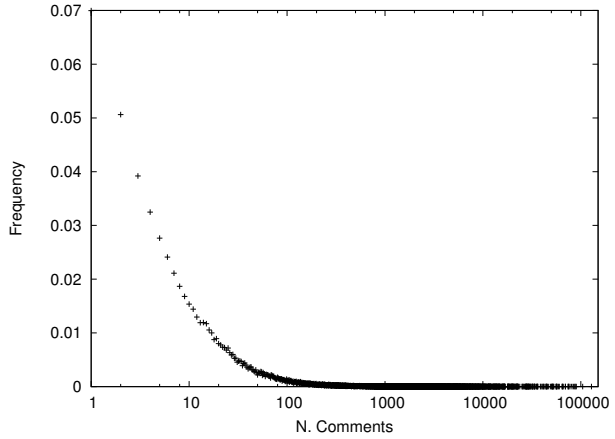

Figure 2: Distribution of Number of Comments per Video

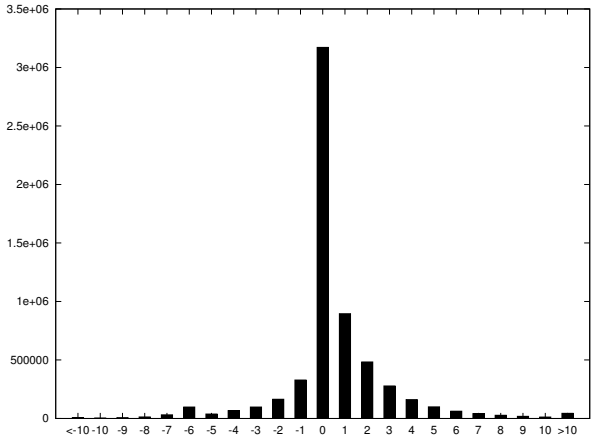

Figure 3: Distribution of comment ratings

Figure 2 shows the distribution of the number of comments per video in the collected set. The distribution follows the expected zipfian pattern, characterized by having most of the energy contained within the first ranked elements as well as subsequent long tail of additional low-represented elements, valid for most community provided data. For our collection, we observe a mean value of $\mu_{comm} = 475$ comments per video, with ratings ranging from $-1,918$ to $4,170$ for a mean value of $\mu_r = 0.61$.

Figure 3 shows the distribution of comment ratings. The following two main observations can be made: On the one hand, the distribution is asymmetric for positive and negative ratings, indicating that the community tends to cast more positive than negative votes. On the other hand, comments with rating 0 represent about 50% of the overall population, indicating that most comments lack votes or are neutrally evaluated by the community.

Preliminary Term Analysis.

The textual content of comments in Web 2.0 infrastructures such as YouTube can provide clues on the community acceptance of comments. This is partly due to the choice of words and language used in different kinds of comments. 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 ranking the terms, we used the Mutual Information (MI) measure [18, 28] 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 in which features and categories ("high community acceptance" and "low community acceptance") are independent of each other.

Table 1: Top-50 terms according to their MI values for accepted (i.e. high comment ratings) vs. not accepted (i.e. low comment ratings) comments

Terms for Accepted Comments				
love	favorit	perfect	wish	sweet
song	her	perform	hilari	jame
best	hot	miss	most	talent
amaz	my	omg	gorgeou	feel
beauti	d	nice	brilliant	avril
awesom	voic	bless	legend	wonder
she	rock	music	ador	janet
thank	lol	sexi	fantast	danc
lt	xd	man	heart	absolut
cute	luv	greatest	time	watch
Terms for Unaccepted Comments				
fuck	ur	game	fuckin	shut
suck	dont	fat	worst	gui
u	ugli	kill	y	im
gai	dick	idiot	pussi	jew
shit	better	dumb	crap	comment
stupid	fag	retard	de	die
bitch	white	bad	cunt	cock
ass	fake	know	bore	name
nigger	black	don	loser	asshol
hate	faggot	sorri	look	read

Table 1 shows the top-50 stemmed terms extracted for each category. Obviously many of the "accepted" comments 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.

4. SENTIMENT ANALYSIS OF RATED COMMENTS

Do comment language and sentiment have an influence on comment ratings? In this section, we will make use of the publicly available SentiWordNet thesaurus to study the connection between sentiment scores obtained from SentiWordNet and the comment rating behavior of the community.

SentiWordNet [9] is a lexical resource built on top of WordNet. WordNet [10] 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. $\{car, automobile\}$).

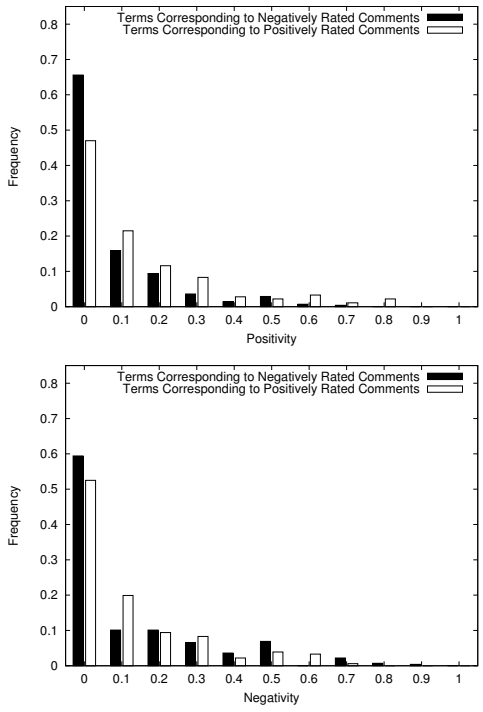

Figure 4: SentiValue histograms for term lists according to MI

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 [8] for an evaluation of these methods). In our experiments, we assign a sentivalue to each comment by computing the averages for *pos*, *neg* and *obj* over all words in the comment that have an entry in SentiWordNet.

A SentiWordNet-based Analysis of Terms.

We want to provide a more quantitative study of the interrelation between terms typically used in comments with high positive or negative ratings. To this end, we selected the top-2000 terms according to the MI measure (see previous section) for positively and negatively rated comments, and retrieved their sentivalue triples (pos, neg, obj) from SentiWordNet if available.

Figure 4 shows the histograms of sentivalues for these terms. In comparison to terms corresponding to positively rated comments, we can observe a clear tendency of the terms corresponding to negatively rated comments towards higher negative sentivalue assignments.

Sentiment Analysis of Ratings.

Now we describe our statistical comparison of the influence of sentiment scores in comment ratings. For our analysis, we restricted ourselves to adjectives as we observed

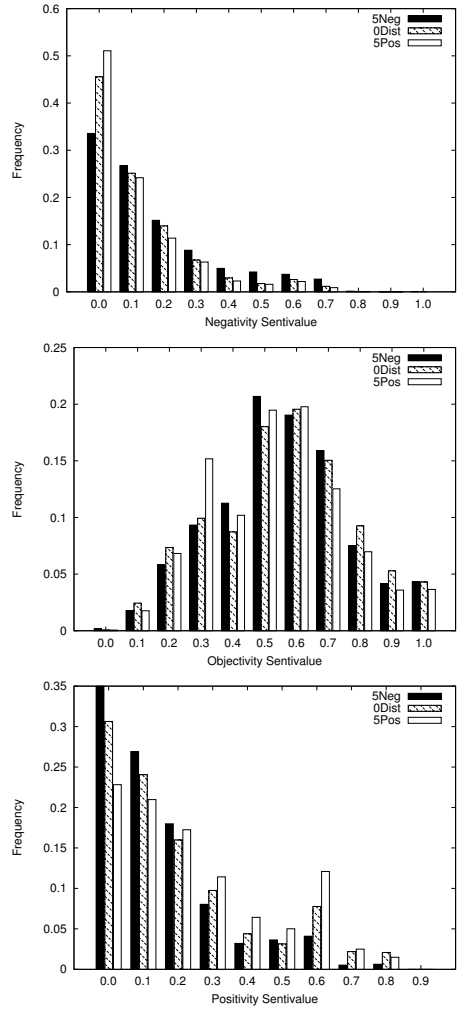

Figure 5: Distribution of comment negativity / objectivity / positivity

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 positive, objective and negative for each different partition. Detailed comparison histograms for these sentiments are shown

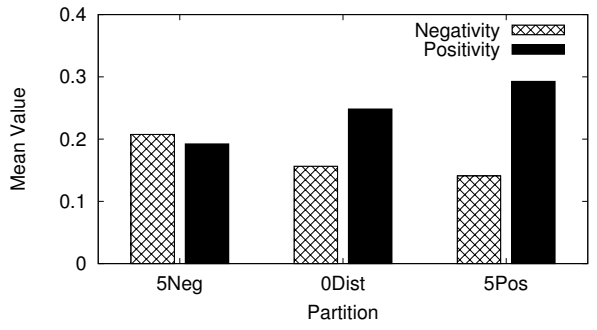

Figure 6: Difference of Mean values for sentiment categories

in Figure 5. These figures provide graphical evidence of the intuition stated above. Negatively rated comments (**5Neg**) tend to contain more negative sentiment terms than positively rated comments (**5Pos**), reflected on 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 line with the initial intuition.

We further analyzed whether the difference of sentivalues across partitions was significant. We considered comment positivity, objectivity and negativity as dependent variables. Rating partition (**5Neg**, **0Dist**, **5Pos**) was used as the independent variable (grouping factor) of our test. Let us denote μ_s^k the mean value for sentiment $s \in \{N, O, P\}$ (negativity, objectivity and positivity respectively) for partition $k \in \{5Neg, 0Dist, 5Pos\}$. Our initial null hypothesis states that the distribution of sentiment values does not depend on the partition states, i.e. the mean value of each independent variable is equal across partitions $H_0 : \mu_s^{5Neg} = \mu_s^{0Dist} = \mu_s^{5Pos}$. The alternative hypothesis H_a states that the difference is significant for at least two partitions. We then used three separate one-way ANOVA (Analysis of Variance) procedures [6], a statistical test of whether the means of several groups are all equal, to verify the null hypothesis, H_0, for each variable negativity (F_N), objectivity (F_O) and positivity (F_P).

We selected a random sample of 15,000 comments. From this, we discarded comments for which sentiment values were unavailable in SentiWordNet, resulting in a final set of 5047 comments. All tests resulted in a strong rejection of the null hypothesis H_0 at significance level 0.01. Figure 6 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.

The ANOVA test does not provide information about the specific mean values μ_s^k that refuted H_0. Many different post-hoc tests exist to reveal this information. We used the Games-Howell [6] test to reveal these inter-partition mean differences because of its tolerance for standard deviation heterogeneity in data sets. For negativity, the following homogeneous groups were found: { {**5Neg**}, {**0Dist**, **5Pos**} }. Finally, for positivity the following homogeneous groups were found: { {**5Neg**}, {**0Dist**}, {**5Pos**} }. These results provide statistical evidence of the intuition that negatively rated comments contain a significantly larger number of negative sentiment terms, and similarly for positively rated comments and positive sentiment terms.

5. PREDICTING COMMENT RATINGS

Can we predict community acceptance? We will use support vector machine classification and term-based representations of comments to automatically categorize comments as likely to obtain a high overall rating or not. Results of a systematic and large-scale evaluation on our YouTube dataset show promising results, and demonstrate the viability of our approach.

5.1 Experimental Setup for Classification

Our term- and SentiWordNet-based analysis in the previous sections indicates that a word-based approach for classification might result in good discriminative performance. In order to classify comments into categories "accepted by the community" or "not accepted", we use a supervised learning paradigm which is based on training items (comments in our case) that need to be provided for each category. Both training and test items, which are later given to the classifier, are represented as multi dimensional feature vectors. These vectors can, for instance, be constructed using tf or $tf \cdot idf$ weights which represent the importance of a term for a document in a specific corpus. Comments labeled as "accepted" or "not accepted" are used to train a classification model, using probabilistic (e.g., Naive Bayes) or discriminative models (e.g., SVMs).

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).

Linear support vector machines (SVMs) construct a hyperplane $\vec{w} \cdot \vec{x} + b = 0$ that separates a set of positive training examples from a set of negative examples with maximum margin. For a new previously unseen comment $\vec{c}$, the SVM merely needs to test whether it lies on the "positive" side or the "negative" side of the separating hyperplane. We used the SVMlight [14] implementation of linear support vector machines (SVMs) with standard parameterization in our experiments, as this has been shown to perform well for various classification tasks (see, e.g.,[7, 13]).

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 = 1000, 10000, 50000,

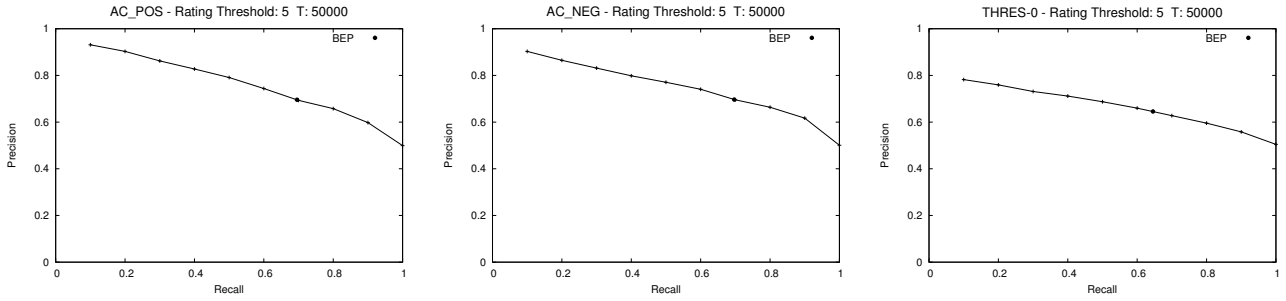

Figure 7: Comment Classification: Precision-recall curves (50000 training comments per class, rating≥5)

Table 2: Comment Classification Results (BEPs)

AC_POS			
T	Rating ≥ 2	Rating ≥ 5	Rating ≥ 7
1000	0.6047	0.6279	0.6522
10000	0.642	0.6714	0.6932
50000	0.6616	0.6957	**0.7208**
200000	0.6753	-	-
AC_NEG			
T	Rating ≥ 2	Rating ≥ 5	Rating ≥ 7
1000	0.6061	0.629	0.6531
10000	0.6431	0.6724	0.6943
50000	0.6627	0.6966	**0.7215**
200000	0.6763	-	-
THRES-0			
T	Rating ≥ 2	Rating ≥ 5	Rating ≥ 7
1000	0.5516	0.5807	0.6014
10000	0.5812	0.6264	0.6424
50000	0.6003	0.6456	0.6639
200000	0.6106	0.6586	**0.6786**

200000) as positive examples and the same amount of randomly chosen "unaccepted" comments as negative samples (where that number of training comments and at least 1000 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 and Conclusions

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 2. The detailed precision-recall curves for the example case of T=50000 training comments class and thresholds +5/-5 for "accepted"/"unaccepted" comments are shown in Figure 7. The main observations are:

- All three types of classifiers provide good performance. 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 0.7208. Consistently, similar observations can be made for all examined configurations.
- Trading recall against precision leads to applicable results. For instance, we obtain prec=0.8598 for recall=0.4, and prec=0.9319 for recall=0.1 for **AC_POS**; this is useful for finding candidates for interesting comments in large comment sets.
- Classification results tend to improve, as expected, with an increasing number of training comments. Furthermore, classification performance increases with higher thresholds for community ratings for which a comment is considered as "accepted".

6. COMMENT RATINGS AND POLARIZING YOUTUBE CONTENT

In this section, we will study the relationship between comment ratings and polarizing content, more specifically tags/topics and videos. By "polarizing content" we mean content likely to trigger diverse opinions and sentiment, examples being content related to the war in Irak 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.

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 8 shows examples of videos with high versus low rating variance (in our specific examples videos about an Iraki girl stoned to death, Obama, and protest on Tiananmen Square in contrast to videos about The Beatles, cartoons, and amateur music). To show the relation between comment ratings and polarizing videos, 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 on a 3-point Likert scale (3: polarizing, 1: rather neutral, 2: in between). The assessments of the

Figure 8: Videos with high (upper row) versus low variance (lower row) of comment ratings

different users were averaged for each video, and we computed the inter-rater agreement using the κ-measure [21], 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).

Variance of Comment Ratings as Indicator for Polarizing Topics.

We were also studying 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 3 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 quantitatively evaluated this tendency in a user experiment with 3 assessors similar to the one described for videos using the same 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 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).

Table 3: 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

7. CATEGORY DEPENDENCIES OF RATINGS

Videos in YouTube belong to a variety of categories such as "News & Politics", "Sports" or "Science". Given that different categories attract different types of users, an interesting question is whether this results in different kinds of comments, discussions and feedback.

7.1 Classification

In order to study the influence of categories on the classification behavior, we conducted a similar experimental series as described in section 5. In the following paragraphs, we describe the results of classification of YouTube comments into the classes "accepted" and "not accepted" as introduced in the previous subsection. In each classification experiment we restricted our training and test sets to comments from the same class. We used smaller training sets than in section 5 as we had less comments available per category than for the overall dataset.

Figure 9 shows the precision-recall curves as well as the break-even-points (BEPs) for comment classification for the configuration T=10,000 training documents and threshold +5/-5 for accepted/unaccepted comments. We observe that training and classifying on different categories leads to clear differences in classification results. While classifiers applied within the categories "Music" and "Entertainment" show comparable performance, the performance drops for for "News & Politics". This might be an indicator for more complex patterns and user relationships for that domain.

7.2 Analysis of comment ratings for different categories

In this section we consider the analysis of comment rating distribution across different categories. Our intuition is that some topics are more prone to generate intense discussions than others. Differences of opinion will normally lead to an increasing number of comments and comment ratings, affecting the distribution.

Figure 10 shows the distribution of comment ratings for a set of selected categories from our subset. We observe several variations for the different categories. For instance, science videos present a majority of 0-scored comments, maybe due to the impartial nature of this category. Politics videos have significantly more negatively rated comments than any other category. Music videos, on the other hand, have a

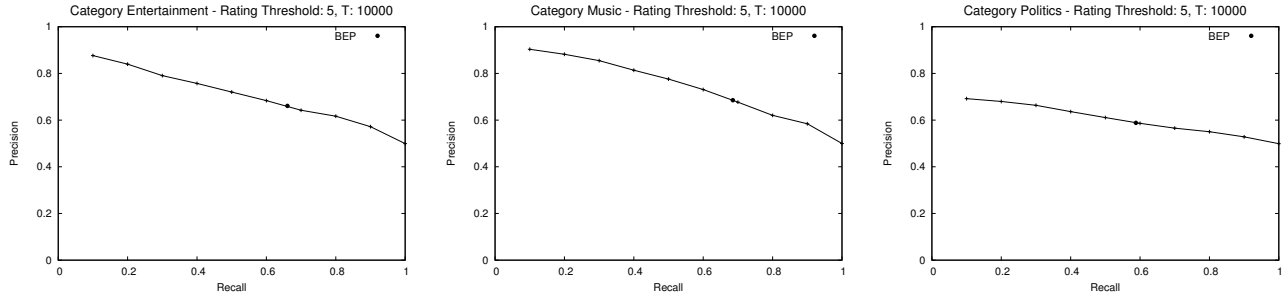

Figure 9: Classification Precision-Recall Curves for Multiple Categories

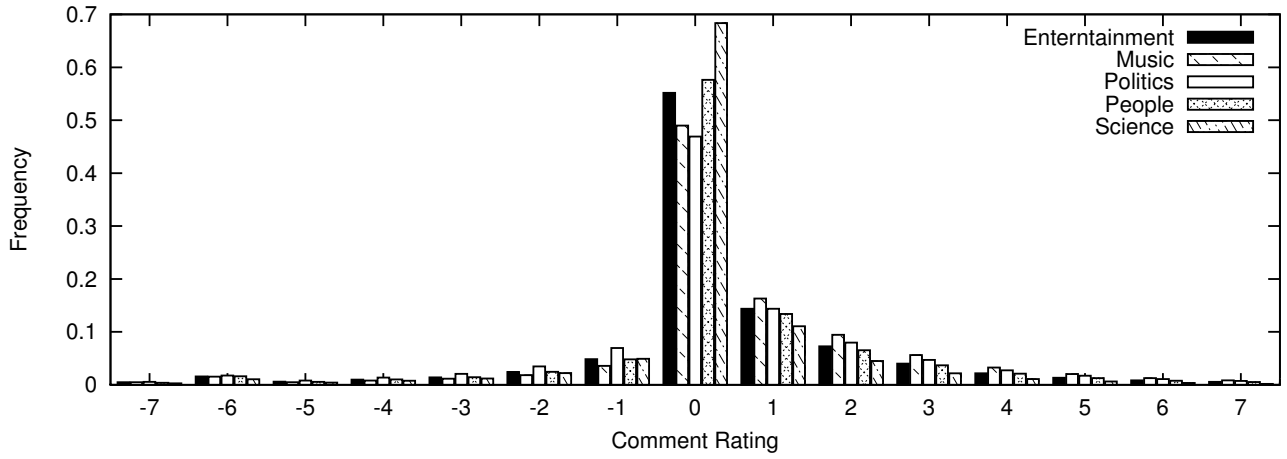

Figure 10: Distribution of comment ratings for different categories

clear majority of positively rated comments. Mean rating score values for all categories in our database are shown in Figure 11.

We further analyzed whether the rating score difference across categories was significant. We considered comment ratings as the dependent variable, and categories as the grouping factor. Let us denote μ_r^i the mean rating score value for category i. We wanted to refute hypothesis H_0 : $\mu_r^i = \mu_r^j, \forall\, i, j$ (i.e. comment ratings mean value is identical for all categories). Our alternative hypothesis H_a states that at least two categories, i and j, feature mean rating scores that are statistically different. We used one-way ANOVA to test the validity of the null hypothesis. For this experiment we considered the complete data set, excluding comments with 0 ratings and no assigned category, for a total of $2,539,142$ comments. The test resulted in a strong rejection of the hypothesis H_0 at significance level 0.01, providing evidence that mean rating values across categories are statistically different.

A subsequent post-hoc Games-Howell test was conducted to study pair-wise differences between categories. Table 4 shows the homogeneous groups found. The table identifies category "Music" as having significantly higher comment ratings than any other, and categories "Autos&Vehicles", "Gaming" and "Science" having significantly lower comment ratings. While some categories are likely to be affected by the lack of comment ratings ("Science"), the significantly lower comment ratings in some categories like "Gaming"

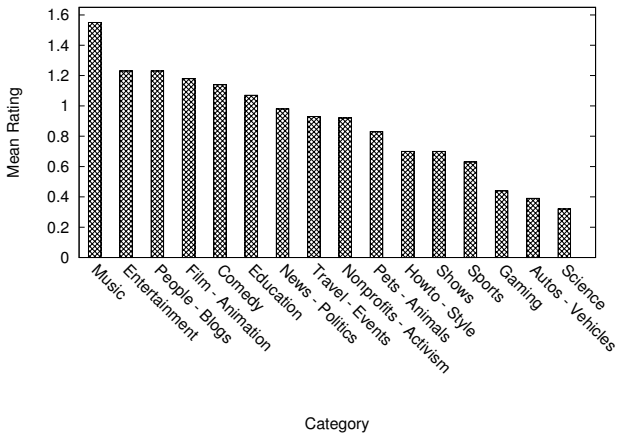

Figure 11: Mean Rating Score per Comment for different Categories

might indicate that malign users (trolls, spammers, ...) are more dominant in these categories than in others.

7.3 Sentivalues in Categories

In Section 4 we provided statistical evidence of the dependency of comment ratings on sentivalues. In this section we extend the analysis to also consider categories, to check whether we can find a dependency of sentivalues for differ-

Table 4: Homogeneous Groups by Mean Rating

	Homogeneous Category Groups
Highest Mean	Music
Medium Mean	Pets&Animals, Comedy, Education Entertainment, News&Politics Nonprofits&Activism, Sports People&Blogs, Shows Travel&Events, Howto&Style
Lowest Mean	Autos&Vehicles, Gaming, Science

ent categories, and provide additional ground to the claims presented in Section 7.2.

We proceeded similarly to Section 7.2. In this case, we considered sentivalue negativity, objectivity and positivity as dependent variables, and categories as the grouping factor. We denote $\mu_r^{N,i}$ the mean negativity value for category i. Analogously, $\mu_r^{O,i}$ and $\mu_r^{P,i}$ denote mean objectivity and positivity values for category i. We wanted to refute hypothesis $H_0 : \mu_r^{K,i} = \mu_r^{K,j}, \forall\ i, j, K \in \{N, O, P\}$ (i.e. comment ratings mean value is identical for all categories). Our alternative hypothesis H_a states that at least two categories, i and j, feature mean values that are statistically different. We used three one-way ANOVA procedures to test the validity of the null hypothesis. For this experiment we considered the complete data set, excluding comments for which sentivalues were not available, for a total of $2,665,483$ comments. The test resulted in a strong rejection of the hypothesis H_0 at significance level 0.01 for the three cases, providing evidence that mean sentivalues across categories are statistically different. Figure 12 shows mean values for sentiments negativity, objectivity and positivity for different categories. Results are in agreement with findings of Section 7.2 (Table 4 and Figure 11). For instance, music exhibits the lowest negativity sentivalue and the highest positivity sentivalue.

Our interpretation of these results is that different categories tend to attract different kinds of users and generate more or less discussion as a function of the controversy of their topics. This clearly goes along with significantly different ratings and sentivalues of comments associated to videos. As a result, user generated comments tend to differ widely across different categories, and therefore the quality of classification models gets affected (illustrated in section 7.1).

8. CONCLUSION AND FUTURE WORK

We conducted an in-depth analysis of YouTube comments to shed some light on different aspects of comment ratings for the YouTube video sharing platform. How does community feedback on comments depends on language and sentiment expressed? Can we learn models for comments and predict comment ratings? Does comment rating behavior depend on topics and categories? Can comment ratings be an indicator for polarizing content? These are some of the questions we examined in this paper by analyzing a sample of more than 6 million YouTube comments and ratings. Large-scale studies using the SentiWordNet thesaurus and YouTube meta data revealed strong dependencies between different kinds of sentiments expressed in comments, comment ratings provided by the community and topic orientation of the discussed video content. In our classification

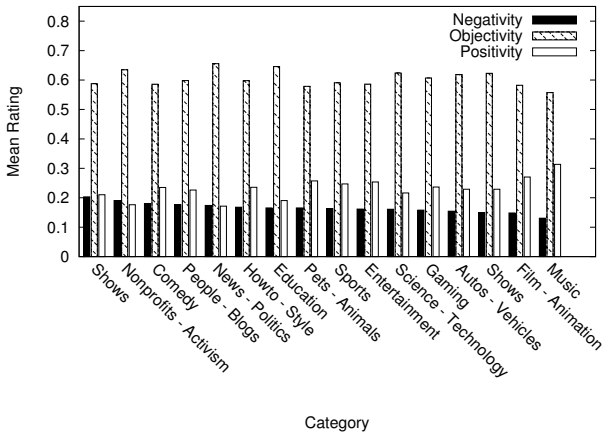

Figure 12: Distribution of comment sentivalues

experiments, we demonstrated that community feedback in social sharing systems in combination with term features in comments can be used for automatically determining the community acceptance of comments. User experiments show that rating behavior can be often connected to polarizing topics and content.

Regarding future work, we plan to study temporal aspects, additional stylistic and linguistic features, relationships between users, and techniques for aggregating information obtained from comments and ratings. We think that temporal aspects such as order and timestamps of comments and upload dates of commented videos can have a strong influence on commenting behavior and comment ratings, and, in combination with other criteria, could help to increase the performance of rating predictors. More advanced linguistic and stylistic features of comment texts might also be useful to build better classification and clustering models. Finally, comments and ratings can lead to further insights on different types of users (helpful users, spammers, trolls, etc.) and on social relationships between users (friendship, rivalry, etc). This could, for instance, be applied for identifying groups of users with similar interest and recommending contacts or groups to users in the system.

We think that the proposed techniques have direct applications to comment search. When searching for additional information in other users' comments, automatically predicted comment ratings could be used as an additional ranking criterion for search results. In this connection, integration and user evaluation within a wider system context and encompassing additional complementary retrieval and mining methods is of high practical importance.

9. ACKNOWLEDGEMENTS

This work was supported by EU FP7 integration projects LivingKnowledge (Contract No. 231126) and GLOCAL (Contract No. 248984) and the Marie Curie IOF project "Mieson".

10. REFERENCES

[1] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *ICML '05: Proceedings*

of the 22nd international conference on Machine learning, pages 89–96, New York, NY, USA, 2005. ACM.

[2] S. Chakrabarti. *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan-Kauffman, 2002.

[3] X. Cheng, C. Dale, and J. Liu. Understanding the characteristics of internet short video sharing: Youtube as a case study. In *Technical Report arXiv:0707.3670v1 cs.NI*, New York, NY, USA, 2007. Cornell University, arXiv e-prints.

[4] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, and L. Lee. 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*, pages 141–150, New York, NY, USA, 2009. ACM.

[5] K. Denecke. Using sentiwordnet for multilingual sentiment analysis. In *Data Engineering Workshop, 2008. ICDEW 2008*, pages 507– 512, 2009.

[6] J. L. Devore. *Probability and Statistics for Engineering and the Sciences*. Thomson Brooks/Cole, 2004.

[7] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *CIKM '98: Proceedings of the seventh international conference on Information and knowledge management*, pages 148–155, Bethesda, Maryland, United States, 1998. ACM Press.

[8] A. Esuli. *Automatic Generation of Lexical Resources for Opinion Mining: Models, Algorithms and Applications*. PhD in Information Engineering, PhD School "Leonardo da Vinci", University of Pisa, 2008.

[9] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LRECÂŠ06)*, pages 417–422, 2006.

[10] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[11] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube traffic characterization: a view from the edge. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 15–28, New York, NY, USA, 2007. ACM.

[12] F. M. Harper, D. Raban, S. Rafaeli, and J. A. Konstan. Predictors of answer quality in online q&a sites. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 865–874, New York, NY, USA, 2008. ACM.

[13] T. Joachims. Text categorization with Support Vector Machines: Learning with many relevant features. *ECML*, 1998.

[14] T. Joachims. Making large-scale support vector machine learning practical. *Advances in kernel methods: support vector learning*, pages 169–184, 1999.

[15] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti. Automatically assessing review helpfulness. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 423–430, Sydney, Australia, July 2006. Association for Computational Linguistics.

[16] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou. Low-quality product review detection in opinion summarization. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 334–342, 2007. Poster paper.

[17] Y. Lu, C. Zhai, and N. Sundaresan. Rated aspect summarization of short comments. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 131–140, New York, NY, USA, 2009. ACM.

[18] C. Manning and H. Schuetze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[19] B. Pang and L. Lee. Thumbs up? sentiment classification using machine learning techniques. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia, PA, USA, 2002.

[20] M. Richardson, A. Prakash, and E. Brill. Beyond pagerank: machine learning for static ranking. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 707–715, New York, NY, USA, 2006. ACM.

[21] A. Rosenberg and E. Binkowski. Augmenting the kappa statistic to determine interannotator reliability for multiply labeled data points. In *HLT-NAACL '04: Proceedings of HLT-NAACL 2004: Short Papers on XX*, pages 77–80, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[22] J. San Pedro and S. Siersdorfer. Ranking and classifying attractiveness of photos in folksonomies. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 771–780, New York, NY, USA, 2009. ACM.

[23] S. Siersdorfer, J. San Pedro, and M. Sanderson. Automatic video tagging using content redundancy. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 395–402, New York, NY, USA, 2009. ACM.

[24] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004.

[25] M. Thomas, B. Pang, and L. Lee. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *EMNLP '06: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 327–335, 2006.

[26] M. Weimer, I. Gurevych, and M. Muehlhaeuser. Automatically assessing the post quality in online discussions on software. In *Companion Volume of the 45rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007.

[27] F. Wu and B. A. Huberman. How public opinion forms. In *Internet and Network Economics, 4th International Workshop, WINE 2008, Shanghai, China*, pages 334–341, 2008.

[28] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, pages 412–420, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.