SenTube: A Corpus for Sentiment Analysis on YouTube Social Media

Olga Uryupina[1], Barbara Plank[2], Aliaksei Severyn[1], Agata Rotondi[1], Alessandro Moschitti[1,3]

[1]Department of Information Engineering and Computer Science, University of Trento,
[2]Center for Language Technology, University of Copenhagen,
[3]Qatar Computing Research Institute
uryupina@gmail.com, bplank@cst.dk, severyn@disi.unitn.it,
rotondiagata@gmail.com, moschitti@disi.unitn.it

Abstract

In this paper we present SenTube – a dataset of user-generated comments on YouTube videos annotated for *information content* and *sentiment polarity*. It contains annotations that allow to develop classifiers for several important NLP tasks: (i) sentiment analysis, (ii) text categorization (relatedness of a comment to video and/or product), (iii) spam detection, and (iv) prediction of comment informativeness. The SenTube corpus favors the development of research on indexing and searching YouTube videos exploiting information derived from comments. The corpus will cover several languages: at the moment, we focus on English and Italian, with Spanish and Dutch parts scheduled for the later stages of the project. For all the languages, we collect videos for the same set of products, thus offering possibilities for multi- and cross-lingual experiments. The paper provides annotation guidelines, corpus statistics and annotator agreement details.

1. Introduction

Social media and streams, such as Twitter, Facebook or YouTube, contain rapidly changing information generated by millions of users that can dramatically affect the reputation of a person or an organization. This raises the importance of automatic extraction of sentiments and opinions expressed in social media. While sentiment analysis for more conventional data has recently attracted a lot of attention from both industry and academia, the paucity of manually annotated data makes these studies only partially useful for social media and streams.

In this paper we present SenTube[1]—a dataset of user-generated comments on YouTube videos annotated for *information content* and *sentiment polarity*. To address the specifics of the YouTube data, we go beyond commonly used per-document sentiment labels: we distinguish between user sentiments expressed with respect to the video and the product discussed at the comment level. This offers valuable annotations for experiments on targeted sentiment analysis, in particular, on online reputation management on the social media and social streams. We also provide data for other NLP tasks, for example, spam detection and document classification on social streams and joint modeling of these phenomena.

It should be noted that, several annotation projects have been proposed recently to develop sentiment analysis models adapted to social media, focusing mainly on Twitter. While the latter provides valuable data for extracting and tracking opinions, the derived corpora are *unstable*: due to the Twitter distribution restrictions, the tweets are only represented with their IDs, without explicit inclusion of their textual content. End users are required to use the Twitter API to download the tweet messages. The substantial amount of tweets either changing or disappearing over time makes impossible a fair comparison of experimental results obtained on these datasets. The SenTube corpus, on the contrary, comes with all the supporting textual content. It can be downloaded and used for NLP experiments in a straightforward way.[2] Our dataset, thus, gives the possibility to work on an important social media context, i.e., comments on YouTube videos. In addition, since the language of YouTube comments and tweets is somewhat similar, we believe that our corpus provides a reliable testbed for sentiment analysis for other types of social media as well, without raising reproducibility issues.

The SenTube corpus will cover several languages. At the moment, it contains English and Italian, with Spanish and Dutch parts scheduled for the later stages of the project. For all the languages, we consider videos for the same set of products, thus offering possibilities for cross-lingual experiments. At the moment, we cover two product domains, *Tablets* and *Automobiles*. For each product, we consider two types of videos: *Technical Reviews* and *Commercials*. Table 1 provides corpus statistics.

The corpus includes not only comments themselves, but also links to the corresponding videos, allowing for joint text and multimedia modeling, building combined models of speech, image, video and text.

We believe that the YouTube comments corpus will provide valuable data to our community for a number of challenging tasks.

2. Related Work

The most commonly used datasets for sentiment analysis include: MPQA corpus of news documents (Wiebe et al., 2005), web customer review data (Hu and Liu, 2004), Amazon review data (Blitzer et al., 2007), JDPA corpus of blogs (Kessler et al., 2010), etc. They contain relatively clean and focused mid-length documents. It must be noted that some of these corpora contain opinion labels induced automatically from user-generated ratings and not labeled manually

[1] http://disi.unitn.it/~haponchyk/ikernels/projects/sentube/

[2] The corpus is currently available on request from the University of Trento and will be made publicly available at the end of the annotation project.

English				
	Tablets		Automobiles	
	videos	comments	videos	comments
reviews	111	18920	29	7670
commercials	28	3153	40	6124
total	139	22073	69	13794
Italian				
reviews	95	5607	51	1758
commercials	5	32	47	2994
total	100	5639	98	4752

Table 1: Corpus statistics: distribution across video types and product domains

by experts.

The aforementioned corpora are, however, only partially suitable for developing models on social media (see Section 3. below). With no gold annotated data available, Pak and Paroubek (2010) present an approach to collecting opinion-mining data from Twitter automatically. A very recent initiative, RepLap (Amigó et al., 2012; Amigó et al., 2013) has been aimed at manually annotating tweets with sentiments towards predefined entities ("reputation"). While RepLab datasets present valuable testbeds for opinion mining on the Social Media data, they are not provided in a standard form for NLP resources due to the Twitter distribution restrictions. Thus, tweets are not represented as textual fragments but only as unique IDs, and potential corpus users are supposed to download their text bodies from Twitter using their own means. This raises an issue with replicability, since tweets tend to disappear with time.

User-generated comments have been successfully used for a variety of NLP and IR tasks, for example, for weblog summarization (Hu et al., 2008) and clustering (Li et al., 2007). Several studies have focused on different aspects of weblog or newspaper reputation/popularity, trying to assess them from commenting patterns (Mishne and Glance, 2006; Rangwala and Jamali, 2010; Park et al., 2011). While most studies focus on weblog comments, several papers aim at analyzing YouTube comments. For instance, Siersdorfer et al. (2010) investigate relations between views, ratings, comments and topics. Yee et al. (2009) incorporate user-generated comments into the search index, showing a significant improvement in search accuracy. In absence of manually labeled data, such algorithms rely on automatically induced approximations of comments' sentiments, for example, Siersdorfer et al. (2010) focus on exploiting user ratings (counts of 'thumbs up/down' as flagged by other users) to YouTube video comments to train classifiers to predict the community acceptance of new comments.

This highlights the importance of a stable corpus of user-generated comments, manually annotated for targeted opinions. Unlike tweets, YouTube comments can be distributed as textual fragments and thus constitute a stable corpus for sentiment analysis on the social media and social streams data.

3. Sentiments on Social Streams

Given the growing demand for social media and streams access, NLP tools that can cope with this new kind of data are becoming more and more important. Social media include, amongst others, Twitter, Youtube, LinkedIn, Facebook. They provide virtual platforms where users communicate in an informal language to share information, news, opinions, etc.

Compared to more conventional text sources, social media and streams pose additional challenges for Information Extraction and Natural Language Processing. Most messages are very short: in some cases, formal requirements apply (e.g., tweets cannot exceed 140 characters), while other social media services support longer messages but most users still generate very short texts. The language of social media is very informal, with numerous accidental and deliberate errors and grammatical inconsistencies: unlike in newswire or weblogs, there is virtually no post-editing and filtering of the incoming textual stream. As a result, social media documents contain numerous out-of-vocabulary words posing challenges for simple bag-of-words NLP models. Another property of the no-editing and no-filtering approach is a very high diversity of the data: a tweet with a specific hashtag or a comment to a specific YouTube video can address many different issues and topics, unlike in review corpora, where the messages are more focused on a target topic.

While a lot of work has been published on Twitter (see Section 2. above), the YouTube data remains only partially covered. Not only YouTube comments exhibit all the typical properties of social media documents discussed above, they also provide an additional interesting dimension: YouTube comments are organized in threads and thus can be used as a testbed for NLP systems modeling user conversations and discussions. Tweets, on the contrary, are often self-contained and do not exhibit an explicit conversational structure.

In the SenTube corpus, we consider two types of videos for each product: technical reviews and commercials. The percentage of commenters who are experts in the subject or at least are well-informed users is superior for technical reviews. Such experts produce elaborated and coherent comments with motivated opinions using proper terminology. Therefore, we expect the information quality as well as vocabulary distributions vary considerably across the two types of videos.

The type of users involved, the quality of information and the language depend also on the topic of the video. Thus, for well-known products, such as iPad or similar popular devices, there is a multitude of comments that are very opinionated, often lack motivation and have a poor information quality.

Another variable influencing commenting patterns is the user who uploads the video. Some users that are experts of a product write technical reviews. Their videos often get comments from other expert users who produce high quality judgements about products, and many video-related comments where the topic is the reviewer itself. This observation highlights the importance of integrating global information (e.g. user profiles) into a sentiment analyzer for the social media data.

Content	
Is the comment related to the product?	yes/now
Is the comment related to the video?	yes/now
Is the comment a spam message?	yes/now
Is the comment written in a language other than English?	yes/now
Is the content off-topic or unclear?	yes/know
Quality of information content	
How much information does the comment provide? How well is it argumented?	0-3 stars
Sentiment Polarity	
Is the comment positive w.r.t. the product?	yes/now
Is the comment negative w.r.t. the product?	yes/now
Is the comment positive w.r.t. the video?	yes/now
Is the comment negative w.r.t. the video?	yes/now

Table 2: Overview of the annotation scheme

4. Annotation guidelines

In the SenTube corpus, we target videos (technical reviews and commercials) featuring commercial products, such as, automobiles, digital cameras or tablets. For each video, we extract user-generated comments and annotate them for the target and the polarity of the expressed opinion, as well as the amount of information explicitly present in the comment. We created the product list in collaboration with the organizers of the RepLab initiative (Amigó et al., 2012; Amigó et al., 2013) for Online Reputation Management on Twitter. Hence, our corpus complements the RepLab dataset, adding another dimension for the reputation analysis.

We extracted the list of videos for each language semi-automatically: after querying the YouTube API for the names of our products and some keywords (e.g., "commercial"), we manually filtered the output to exclude irrelevant videos. We downloaded all available comments of each video, through the YouTube API. The comments are listed in chronological order (the oldest comments first). This gives the annotators some context that might be necessary for the correct interpretation of a specific comment.[3]

We have created a web-based annotation tool that allows for fast and efficient annotation (cf. Figure 1). The annotator is provided with the video and the ordered stream of comments. After watching the video, the annotator is supposed to process the comments one by one, assigning comment-level labels. The annotation is organized hierarchically to save manual effort and enforce data consistency: for example, the polarity can only be marked for *product-* and *video-related* comments. Table 2 shows a list of questions to be answered by our annotators for each comment. The annotation guidelines are summarized below.

Product relatedness. Comments that discuss the topic product in general or some features of the product are considered *product-related*. Possible features include:

- internal properties: colour, weight, technical specifications
- external properties: price, specifics of the delivery, availability in different countries, perception/image, software and so on. on: "what's the name of the wallpaper?"
- sentiments: "I want this! I'm gonna buy it!"
- comparing/contrasting to other related products. Typically, such comments express non-ambiguous sentiment, otherwise they are marked as *off-topic*.

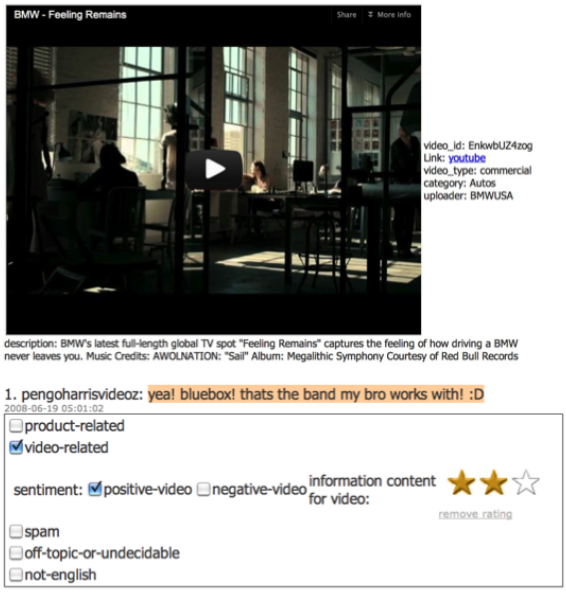

Figure 1: The YouTube annotation tool.

Such comments are annotated as product-related, even if the features are discussed in an implicit way: "Like I said.. her WiFi probably reaches outside" – this comment implies that the product has WiFi and thus is to be considered product-related, even though it does not mention the product itself (`Kindle Fire`).

Comments that discuss alternative products are annotated as product-related even if they do not mention the topic product explicitly: "Oh, you mean an iPhone" is a product-related comment even if the topic product is `Kindle Fire`, not `iPhone`.

Video relatedness. Comments that discuss the video or some of its details are annotated as *video-related*:

[3]The API does not give access to all the comments from the date the video was originally posted and thus even the very first comments can be out of context.

- discussing video in general, expressing positive/negative sentiment: "Great review!";
- requesting/providing information on the crew, location, soundtrack and so on; and
- requesting/providing information on other products seen in the video that are not alternatives to the topic product (e.g., clothing of the main character).

Spam. Comments that provide advertising and malicious links are annotated as *spam*. Such comments might contain:

- bare links
- links and illegible (automatically generated) text
- links and well-formed text, often it's related to the product: "OMG! I got an ipad from this site: getipad2.us.mn - - I was skeptical but it works!"
- suspicious instructions, suggesting googling technical terms, entering keywords in some applications, submitting passwords etc.

Only malicious links are considered *spam*. Thus, if a user provides a link to the webpage of the video crew or its soundtrack, the comment is annotated as *video-related*, not as *spam*.
In particular, links provided by the author of the video in response to various information requests, are not annotated as *spam*.
Spam comments are not labeled for product- or video-relatedness, sentiment polarity or information quality
Non-English. Comments that are not written in English are annotated as *non-English*. They are not labeled for product- or video-relatedness, sentiment polarity or information quality, even if the content is clear (e.g., "Preferisco iPhone" is not a *product-related* comment and receives 0 stars for information quality): "w Polsce od razu kto by jej ukrad t paczk spod drzwi xD". The following types of comments are not to be tagged as *non-English*:

- slang, especially Internet slang ("lmao")
- texting and other types of deliberate misspelling ("cu", "dos leg")
- very bad English
- comments containing virtually only digits and similar characters ("0:15", "10mp", "+1", ";P").

Similarly, we annotate **non-Italian**, **non-Spanish** and **non-Dutch** comments in the respective parts of the corpus.
Off-topic/unclear. Comments that have very little content ("lmao") or content that is not related to the video ("Thank you!") should be annotated as *off-topic/unclear*.
Information quality. The annotators score comments with respect to the amount/quality/specificity of the information they contain: depending on the quality of the comment, the *information content* can be assigned from 0 to 3 stars. We annotate separately the quality of the information with respect to the product and the video.

- *0 stars* comments contain no information with respect to the video or the product. These include *spam*, *not-English* and *offtopic/unclear* comments: "Yayyy!!!", "Thanks buddy!", "Dud, show a bit more respect. will you?", ":P".
- *1 star* comments contain some information, but it's either very generic or expressed in an extremely simplistic form: "5GB" (1 star w.r.t. the product), "Mia Sara" [name of the actress] (1 star w.r.t. the video), "Cool I want the ipad :)" (1 star w.r.t. the product), "Getting one today!!!" (1 star w.r.t. the product). Questions related to the product that have no presuppositions are also annotated with 1 star: "what's the name of the wallpaper?" (1 star w.r.t. the video), "Do we have 4G in this country??" (1 star w.r.t. the product).
- *2 stars* comments contain contain some specific information, but it's only partially argumented and/or motivated: "I hace one of this, battery sucks" (2 stars w.r.t. to the product) "It's 50grams heavier" (2 stars w.r.t. to the product) Questions that make assertions about specific properties of the product/video are annotated with 2 stars: "Great review!!!! What bumper do you have on your Iphone????" (1 star w.r.t. the video, 2 stars w.r.t. to the product), "I heard that it takes forever to charge the ipad 3 is that true?" (2 stars w.r.t. to the product)
- *3 stars* comments are well argumented (for example, contain comparisons, lots of information about video or product aspects, may also additionally contain a question): "The iPad 2's battery lasts longer, and is supposedly a bit more faster. The new iPad has Retina Display, as in a lot of pixels in that small little device. =) Now what do you want to use the iPad for?" (3 stars w.r.t. to the product)

Sentiments and polarity. We annotate the polarity (positive/negative) of the comment with respect to the product and the video:

- *positive-product* comments express positive sentiments regarding some product aspect or the product in general: "For some reason, I feel this one is going to Rock. Love Asus!!"
- *negative-product* comments express negative sentiments regarding some product aspect: "You'll never get expandable memory, it's not how Apple does things. The front facing camera should have definitely been improved though. VGA quality, on a £400 - £800 tablet, in 2012 is ridiculous! :("
- *positive-video* comments express positive sentiments regarding some video features or the video in general: "brandon your the best on youtube liv your vids", "I know now, she's Karme Boixadera, a model from Spain. Oh hell she's HOT!"
- *negative-video* comments express negative sentiments regarding some video features or the video in general

If the comment contains several statements of different polarities, it is annotated as both *positive* and *negative*:

annotation	# comments	%
Content		
product-related	18790	52.4
video-related	10349	28.8
spam	1055	2.9
non-english	544	1.5
off-topic	7267	20.2
Sentiment Polarity		
+product	5326	14.8
incl. (+/-)product	991	2.8
-product	4925	13.7
+video	3930	10.9
incl. +-video	204	0.6
incl. +video,+product	417	1.2
incl. +video,-product	215	0.6
-video	1577	4.4
incl. -video,+product	149	0.4
incl. -video,-product	77	0.2

Table 3: Comment distribution across categories, English.

annotation	# comments	%
Content		
product-related	5594	53.8
video-related	2832	27.2
spam	28	0.2
non-italian	259	2.5
off-topic	2188	21.1
Sentiment Polarity		
+product	1545	14.9
incl. (+/-)product	320	3.1
-product	1706	16.4
+video	1067	10.3
incl. +-video	81	0.7
incl. +video,+product	157	1.5
incl. +video,-product	68	0.6
-video	520	5.0
incl. -video,+product	45	0.4
incl. -video,-product	66	0.6

Table 4: Comment distribution across categories, Italian

"Love the video but waiting for iPad 4" (*positive-video* and *negative-product*).

If neither *positive* nor *negative* is selected (e.g., if the user discusses some properties of the product/video without giving implicit or explicit judgment), the comment is assumed to be **neutral**: "I'm watching oplyics and they played this commercial".

Some comments do not mention the target product, but discuss the alternatives: "IPads has 100mp cameras" (when discussing other tablets). They are considered *negative* or *positive*, if they assumes that the alternative product is better or worse respectively. Comments that contain strong emotions towards other users ("You're stupid!", "The guy 3 spaces before this comment is an idiot") are not annotated for sentiments, as the emotions are not directed towards any aspects of the product or the video.

Content	
product-related	0.79
video-related	0.75
spam	0.94
not-english	-
off-topic	0.56
Polarity	
positive-product	0.66
negative-product	0.63
positive-video	0.73
negative-video	0.09
Information Quality	
with respect to product	0.64
with respect to video	0.67

Table 5: Annotation agreement for English: α values

5. Corpus characteristics

At the current stage of the project, we have annotated videos in English (35887 comments, 208 videos) and Italian (10391 comments, 198 videos), corresponding to 35 distinct products. The amount of data per video varies considerably, ranging from 4 up to 987 comments per video.

Tables 3 and 4 show the data distribution across annotated categories. As the tables suggest, a considerable number of comments contain opinionated statements on either the video itself or the product discussed/advertised. At the same time, many comments contain opinions on both the video and the product. The same comment sometimes expresses both positive and negative sentiments. These statistics confirm our hypothesis that commonly used annotation schemes with a single opinion label per document might be too coarse-grained here.

At the pilot stage of the annotation project, we asked four annotators to label a sample set of one hundred comments and measured the agreement. To assess the annotator agreement, we use the α value (Krippendorf, 2004; Artstein and Poesio, 2008), reported in Table 5. The α value is an universal measure of the annotator agreement, which is applicable to experiments with more than two coders as well as to annotation schemes with non-binary values (as in our "Information Quality" case).

Once we produced the final version of the annotation guidelines, we assigned the entire annotation task to a single coder, who is annotating all the documents. The current coder did not participate to the reported assessment above. However, we measured the agreement with her using the gold-standard sample, adjudicated by the four annotators, to ensure the annotation quality. With the only exception of the negative-video category, we have achieved a reliable α score, ranging from 60 to 80%.

With any annotated corpora, it is important to provide the performance of baseline models. Our baseline is a bag-of-words model, which is a standard baseline in the sentiment and text classification tasks. We include the results for the tasks of sentiment and comment type classification for videos from two product categories: automobiles and tablets. Our classifier is an SVM with a linear kernel. Table 6 reports averaged accuracies of a multi-class classifier

Task\Domain	Automobiles	Tablets
sentiment	60.6	72.1
comment type	64.1	79.3

Table 6: Accuracies of the bag-of-words models on two tasks: sentiment and comment type classification for English. Two product domains: automobiles (13k) and tablets (21k).

on a 10-fold cross-validation experiment.

6. Conclusion

In this paper we have presented the SenTube corpus—an annotation project at the University of Trento aiming at creating a public benchmark for text categorization and targeted opinion mining of user-generated comments on YouTube videos. The dataset contains comments to two distinct types of videos: technical reviews and commercials. It covers products from different domains (automobiles, tablets, digital cameras), thus providing possibilities for domain adaptation studies. At the present stage, the corpus covers English and Italian videos on the same products, with Spanish and Dutch parts to be added in the future, allowing for multi- and cross-lingual experiments.

7. Acknowledgments

The research described in this paper has been partially supported by a Google Faculty Award 2011, the Google Europe Fellowship Award 2013 and the European Community's Seventh Framework Programme (FP7/2007-2013) under the grant #288024: LIMOSINE – Linguistically Motivated Semantic aggregation engiNes.

8. References

Enrique Amigó, Adolfo Corujo, Julio Gonzalo, Edgar Meij, and Maarten de Rijke. 2012. Overview of RepLab 2012: Evaluating online reputation management systems. In *CLEF (Online Working Notes/Labs/Workshop)*.

Enrique Amigó, Jorge Carrillo de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Tamara Martín, Edgar Meij, Maarten de Rijke, and Damiano Spina. 2013. Overview of RepLab 2013: Evaluating online reputation monitoring systems. In *CLEF (Online Working Notes/Labs/Workshop)*.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, December.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL*.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *Proceedings of the 19th national conference on Artifical intelligence*, AAAI'04, pages 755–760. AAAI Press.

M. Hu, A. Sun, and E.-P. Lim. 2008. Comments-oriented document summarization: understanding documents with readers' feedback. In *Proceedings of SIGIR*.

Jason S. Kessler, Miriam Eckert, Lyndsie Clark, and Nicolas Nicolov. 2010. The 2010 icwsm jdpa sentment corpus for the automotive domain. In *4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC 2010)*.

Klaus Krippendorf, 2004. *Content Analysis: An Introduction to Its Methodology, second edition*, chapter 11. Sage, Thousand Oaks, CA.

B. Li, S. S. Xu, and Zhang J. 2007. Enhancing clustering blog documents by utilizing author/reader comments. In *Proceedings of ACMSE*.

G. Mishne and N Glance. 2006. Leave a reply: An analysis of weblog comments. In *Proceedings of WWW*.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of LREC*.

S. Park, M. Ko, J. Kim, Y. Liu, and Song J. 2011. The politics of comments: Predicting political orientation of news stories with commenters sentiment patterns. In *Proceedings of CSCW*.

H. Rangwala and S. Jamali. 2010. Co-participation networks using comment information. In *Proceedings of ICWSM*.

S. Siersdorfer, S. Chelaru, W. Nejdl, and J.S. Pedro. 2010. How useful are your comments? analyzing and predicting YouTube comments and comment ratings. In *Proceedings of WWW-2010*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

W.G. Yee, A. Yates, S. Liu, and Frieder O. 2009. Are web user comments useful for search? In *Proceeding of LSDS-IR*.