Noname manuscript No.
(will be inserted by the editor)

Sifting Useful Comments from Flickr Commons and YouTube

Elaheh Momeni · Bernhard Haslhofer ·
Ke Tao · Geert-Jan Houben

the date of receipt and acceptance should be inserted later

Abstract Cultural institutions are increasingly contributing content to social media platforms to raise awareness and promote use of their collections. Furthermore, they are often the recipients of user comments containing information that may be incorporated in their catalog records. However, not all user-generated comments can be used for the purpose of enriching metadata records. Judging the usefulness of a large number of user comments is a labor-intensive task. Accordingly, our aim is to provide automated support for curation of potentially useful social media comments on digital objects. In this paper, the notion of usefulness is examined in the context of social media comments and compared from the perspective of both end-users and expert users. A machine-learning approach is then introduced to automatically classify comments according to their usefulness. This approach uses syntactic and semantic comment features while taking user context into consideration. We present the results of an experiment we conducted on user comments collected from Flickr Commons collections and YouTube. A study is then carried out on

Elaheh Momeni
University of Vienna
Faculty of Computer Science
Tel: +43-1-42-77-78825
Fax: +43-1-4277-39649
E-mail: elaheh.momeni.roochi@univie.ac.at

Bernhard Haslhofer
Austrian Institute of Technology
E-mail: bernhard.haslhofer@gmail.com

Ke Tao
Delft University of Technology
Department of Software and Computer Technology
E-mail: k.tao@tudelft.nl

Geert-Jan Houben
Delft University of Technology
Department of Software and Computer Technology
E-mail: g.j.p.m.houben@tudelft.nl

the correlation between the commenting culture of a platform (YouTube and Flickr) with usefulness prediction. Our findings indicate that a few relatively straight forward features can be used for inferring useful comments. However, the influence of features on usefulness classification may vary according to the commenting cultures of platforms.

Keywords User-generated Comment · Social Media · Usefulness · Prediction · YouTube · Flickr

1 Introduction

Social media sites are gaining increasing importance in cultural institutions for the dissemination of digitized cultural contents: the Library of Congress, for instance, has published more than 18,000 photos organized in 24 sets on Flickr Commons[1]. The British Library maintains several Facebook pages exposing digitized images, manuscripts, and other digital resources. Another example is Tate which disseminates contents on Facebook, Twitter, Flickr and other social media sites[2]. Each support some kind of annotation feature, ranging from simple button-like clicks to user-contributed full-text comments.

Comments can add supplemental information to existing digital resources, which might be interesting for other users. Besides subjective utterances, they may also contain factual information such as names and places depicted on digital media objects, which is not available in existing metadata records. This information can be gathered by institutions to enrich existing descriptive metadata records and then later to support efficient information retrieval and digital resource management [20,12]. For example, a photo that was published on Flickr by the Library of Congress was initially labeled as “Reid Funeral”. A Flickr user added the comment “Photo shows the crowd gathered outside of the Cathedral of St. John the Divine during New York City funeral of Whitewall Reid, American Ambassador to Great Britain.” This comment contains factual information that clearly goes beyond the initial label.[3]

Yet, not all user-generated comments are useful for the purpose of enhancing metadata records due to users having different backgrounds, levels of expertise, and intentions for contributing comments. Consequently, the quality of user-generated comments ranges from very useful to entirely useless; comments can even be abusive or off-topic. And, as may be expected, what makes a USEFUL comment useful is contingent upon a number of factors including the media type (e.g., document, video, art object, photo), the entity type of the object (e.g., is the object associated with a person, place, event), the time period associated with the object (e.g., early 20th century vs. the 1960’s), and even how controversial the object may be. Another factor as to

[1] Library of Congress Flickr Pilot Project Report Summary `http://www.loc.gov/rr/print/flickr_report_final_summary.pdf`

[2] `http://www.tate.org.uk/about/our-work/digital/social-media-directory`

[3] Source: Library of Congress Flickr Pilot Project Report Summary, `http://www.loc.gov/rr/print/flickr_report_final_summary.pdf`.

whether a comment may be useful or not is whether usefulness is judged from the perspective of an institution, which might require objective and informative descriptive annotations, or from the perspective of an end-user, who might value longer, more personal, or more subjective descriptions.

Using a dedicated human curator or forum administrator to moderate social media comments is expensive, time-consuming and often not feasible given the potentially high number of comments and typically small number of staff members in cultural institutions. Accordingly, automated filtering approaches are needed to segregate useful comments from non-useful ones. This means that methods for estimating the usefulness of user-generated comments are gaining increasing attention [26,8,4]. The most common approach simply enables all users to vote on (and possibly moderate) the contributions of others [26,27, 29], thus avoiding an explicit definition of "useful". Nevertheless, Liu et al. [17] show that voting is influenced by a number of factors (e.g., a "rich get richer" phenomena) that distort accuracy.

The goal of the work reported here is to provide *automated* support for the curation of useful user-generated comments for use as descriptive annotations for digital media objects. To this end, the central contributions of this paper can be summarized as follows:

- *Identification of the characteristics of useful comments:* we study two types of digital objects — images and videos — from two popular social media platforms — Flickr Commons[4] and YouTube respectively, and collect users' and experts' usefulness judgements (by using a crowd-sourcing approach) to identify the usefulness of comments gathered. We then identify technical features that can be derived from textual content and the author's context and characterize the usefulness of a comment.
- *Providing an automated method for identifying potentially useful comments.* We apply the technical features in a series of experiments to build a classifier that can automatically identify the usefulness of comments. Furthermore, we investigate to what extent certain topics of media objects play a role with regard to usefulness classification.
- *Study the correlation between the commenting culture of a platform with usefulness prediction.* We investigate to what extent the commenting culture of a platform plays a role with regard to usefulness classification.

We investigate usefulness from the users' perspective, defining a comment as USEFUL if it provides descriptive information about the object beyond the usually very short title accompanying it. With this definition in hand, we employ crowd-sourcing techniques to create a gold standard data set of USEFUL and NON-USEFUL comments and propose the use of standard supervised

[4] "The key goals of The Commons on Flickr are to firstly show users hidden treasures in the world's public photography archives, and, secondly, to show how users' input and knowledge can help make these collections even richer. Users are invited to help describe the photographs they discover in The Commons on Flickr, either by adding tags or leaving comments." `www.flickr.com/commons`

machine learning techniques to develop a "usefulness" classifier that distinguishes useful from non-useful user-generated comments. We consider over thirty features for the classifier including features for readability, informativeness/novelty, syntactic traits, named entity presence, sentiment, topical traits of the text, and features that describe the author's posting and social media behavior.

The examples below show some comments judged as USEFUL or NON-USEFUL by human coders within our experiments.

- USEFUL**: Flickr photo - Dr. F.A. Cook**[5]. "This must be Dr. Frederick A. Cook (1865-1940), the American explorer who claimed to have reached the North Pole in 1908, before Robert Peary. The controversy over his claim continues. Not only does he have a Wikipedia article, but there are websites dedicated both to disdaining him and to celebrating him. Old controversies never die; they just go on the Internet."
- NON-USEFUL**: Flickr photo - Capt. and crew of MACKAY-BENNETT**[6]. " My great grandfather was an engineer at that time. I'd love to get a list of the names in that photo."
- USEFUL**: YouTube video - Lady diana interview before wedding**[7]. "She had JUST turned 20 years old when they married-in fact it had been less than a month since her 20th birthday. She wasn't anything more than a teenager. So tell me- how good were you at judging character at that age eh?"
- NON-USEFUL**: YouTube video- World War I: Battle Of Verdun**[8]. "Rich people get their poor people to fight the other rich people's poor people. And the[n] we do it all over again. Humanity is truly retarded."

Our findings can be summarized as follows: first, we find that our trained classifier identifies useful comments for Flickr photos with high reliability (precision of 0.87 and recall of 0.90) and which outperforms a strong baseline (precision of 65, recall of 80). Although, the identification of useful comments on YouTube proves to be more difficult (precision of 65, recall of 83). Again the classifier outperforms the baseline (precision of 55, recall of 70) [4].

Furthermore, according to our findings, when inferring the usefulness of comments attached to digital media objects, only a few relatively straightforward features can be used. However, having analyzed the importance of features in different topic areas (place, person, and event), it becomes clear that when inferring the usefulness of comments, the influence of features varies slightly depending on topic areas. Psychological content characteristics appear to be the most influential ones. Therefore, being able to determine the topic area of a media object prior to inferring usefulness helps to classify useful comments more accurately [3] and [4].

Analysis of the top-ranked features of the classifier indicates that semantic and topic-based features are very important for accurate classification for both Flickr and YouTube, especially for those that capture subjective tone, sentiment polarity and the existence of named entities. In particular, comments that mention named entities are more likely to be considered USEFUL;

[5] http://www.flickr.com/photos/library_of_congress/2850357813/comment72157607279573241

[6] http://www.flickr.com/photos/library_of_congress/2536790306/comment72157629444651496

[7] http://www.youtube.com/watch?v=Yka3M4uvUyo

[8] http://www.youtube.com/watch?v=d2qamDMs-3g

those that express the emotional and affective processes of the author are more likely to be considered NON-USEFUL. Similarly, terms indicating INSIGHT (e.g., think, know, consider) are associated with USEFULness while those indicating CERTAINTY (e.g., always, never) are associated with NON-USEFUL comments [4].

Next, we discover that performance varies according to the platform's commenting culture. Investigating two different social media platforms — YouTube and Flickr — we find that the classifier is more easily able to recognize useful comments for Flickr. Furthermore, how influential features impact on the usefulness of a comment varies slightly according to the commenting culture of the platform. Thus, to achieve a more accurate classification of useful comments, a model should be trained that takes into account the commenting culture of the platform.

Although we include some features specific to the Flickr and Youtube platforms, we observe how most of them exhibit more generic properties and can also be applied to other platforms. Furthermore, we select two different media objects with three different topics from real-word events, places, and persons from different times. This results in the possibility for a wider application of the proposed approach which is adaptive and usable for other application domains such as news articles. Although other types of media objects could add more influential features for achieving higher accuracy, we also demonstrate with a minimum set of features (which can be extracted from any type of media) that the approach is able to identify the usefulness of comments. Moreover, it is important to note that this work particularly focuses on historical events, persons, and happenings which have had impact in history.

We believe that the findings reported in this article provide the basis for the next steps, which include the implementation of solutions that support content curators in cultural institutions when filtering potentially useful comments from large scale social media datasets. Factual information contained in such comments could be used to create new or enhanced existing metadata descriptions and to subsequently improve content retrieval. However, these steps are beyond the scope of this paper.

This article extends our previously published works in [3] and [4] by a more detailed comparison of usefulness characteristics of user-generated comments from different platforms and by a deeper understanding of the correlation between the commenting culture of a platform with usefulness prediction. This article also gives an overview of methods and techniques we proposed in [3] and [4] for identification and prediction of useful comments in various social media platforms. It is organized as follows: In Section 2 we discuss the notion of usefulness and identify possible characteristics of useful social media comments by analyzing related work on assessing and modeling the quality of user-generated content. Section 3 provides an overview of different technical features to characterize the comment. Section 4 describes our data acquisition process to collect usefulness judgements. Section 5 presents a series of usefulness classification experiments and evaluation of the derived features and provides an interpretation of to what extent the commenting culture of a platform influences the performance of the usefulness classifier. Finally, we discuss and conclude our work in Section 5.

2 Background and Related Work

The Oxford dictionary defines *usefulness* as "*a quality or fact of being able to be used for a particular or in several ways*". Accordingly, when characterizing usefulness the institutional context and the application domain are indispensable factors.

User-generated content, a relatively general term, can include different "*application domains*" such as tags, product reviews, postings in the questions and answers (Q&A) platforms and comments on digital resources and other media. But, each type of user-generated content has different characteristics. For example, useful tags also provide descriptive information for objects despite user-generated comments having different characteristics from user-generated tags. These characteristics may be that the comments are longer and more informal with regard to structure as a result of authors being able to converse, express subjective opinions and emotions, and describe informative useful information about a media resource. Furthermore, the most common approach for assessing and ranking different application domains — such as helpfulness of product review or quality of questions and answers in Q&A platforms — simply allows all users to vote on (and possibly moderate) the contributions of others [26, 27, 29]. However, using machine-based approaches based on crowd-judgments to train a model for identifying high quality content avoids an explicit definition of "usefulness". The goal of the work reported here is to provide a novel, alternative, and *automated* support for the curation of useful user-generated comments. These can be used as descriptive annotations for digital objects, taking into consideration the explicit definition of useful. We have discovered that the following main research contexts have previously discussed the notion of usefulness:

Assessing the quality of questions and answers. Agichtein et al. [1] by combining features from different sources of information propose a general graph-based classification approach for assessing high-quality questions and answers in Q&A platforms. Liu et al. [18] investigate a method for predicting information seeker satisfaction in Q&A platforms and explore a variety of content, structure, and community-focused features for this task. Harper et al. [13] propose an algorithm that classifies questions as informational or conversational.

Assessing the quality of postings in micro-blogging services. Castillo et al. [6] propose automatic methods to assess the quality and credibility of a given set of tweets, first by analyzing postings related to trending topics, and then by classifying them as credible or non-credible. Diakopoulos et al. [8] by using a human centered design approach propose methods for assessing and classifying the variety of sources found through social media by journalists. Becker et al. [2] propose an approach for relevant Twitter selection. They show that the centroid (as a centrality-based approach) method as the most accurate way to select relevant tweets.

Assessing the helpfulness of product reviews. Predicting the helpfulness of a product review (e.g., how many people have considered a particular product review helpful) is another related problems to usefulness. Several approaches show that a few relatively straightforward features can be used to predict with high accuracy, whether a review will be deemed helpful or not. These features are mixture of subjective and objective information [9], length of the review [16,9], checking the number of spelling errors — readability [9], and conformity ("a review is evaluated as more helpful when its star rating is closer to the consensus star rating for the product" [16,7]). Moreover, Lu et al [19] demonstrate how the social features of reviewers can help the classification process.

Assessing the usefulness of user-generated tags. Several works in tagging and folksonomy research discuss the selection of tags that permit people to better describe their content or the assessment of user-generated tags. Assessment of tag co-occurrence patterns are proposed by Sigurbjoernsson and van Zwol [21] for the filtering of useful tags. They show that the tag frequency distribution follows a perfect power law, indicating that the mid section of this distribution contains the most interesting candidates for tag recommendation. Weinberger et al [24] define a metric of tag ambiguity, based on a weighted Kullback-Leibler (KL) divergence of tag distributions. Hall and Zarro [11] explore a comparison of the abstracting and indexing practices of a semi-expert community metadata and the social tags generated by Delicious.com users for the same corpus of materials and show these two groups still remain dissimilar to provide description for the object.

2.1 Taxonomy of Useful Social Media Comments

A preliminary exploration of these assessment and ranking methods demonstrates that some relatively straightforward features and strategies, derived from content and context of comments, can be used to characterize with high accuracy whether a user-generated content (tags, Q&A postings, tweets, and product reviews) is helpful, relevant, high quality, or credible [3]. Table 1 shows a categorization of these featues and strategies into three feature groups.

Although our task is different, we will rely on some of these features for the learning-based classifier. In our work, we define a comment as useful, if it provides additional descriptive information of media objects. More precisely, this paper focuses on understanding the characteristics of useful comments from users as well as experts perspectives and furthermore on the development of automated mechanisms for classifying useful and non-useful comments. We evaluated to what extent users' and experts' perspective of usefulness match. In systems with numerous users and comments automated mechanisms for classifying useful and non-useful comments can support curators and system managers in selecting potentially useful comments and saving time and costs.

Features Groups	Ref	Short Description
Text statistics and syntactic features	[1,6,16,8,9,19, 7]	Aggregate statistics extracted from the text such as length, readability, #token, etc
Semantic and topical features	[8,9,23,24,21, 15,6,17]	The semantics of a comment and its semantic similarity or diversity to other comments, such as subjectivity tone and topical conformity to other comments.
User and social features	[19,18,23,6,8,1]	Different characteristics of users and their social context, #uploaded object, and #contact

Table 1 Abstract overview of features used in related work for characterizing user-generated content [3].

3 Features Engineering

Given the available approaches and features for similar problems, explored in Section 2, we can conclude that straightforward features derived from social media and textual content have been used to accurately characterize whether user-generated content is helpful, relevant, of high quality, or even credible. Therefore, we believe that the features related to the usefulness problem can be constructed with proper hypotheses. Moreover, we have looked into the examples found in the real data set and proposed observable features that are possibly related to the usefulness of the comment.

In the rest of this section, we provide an overview of the different features we use to analyze and estimate the usefulness of a comment. Inspired by the cases we found, all these features are aligned with our assumption of characteristics of useful comments. Although we introduce these features by inspiration we got from the Flickr and the YouTube platforms, most of them are quite generic and can also be applied to other platforms. In Table 2, we list each feature along with a short description. We grouped these potentially important features into three different groups according to the feature categorization we introduced in Section 2.

Text Statistics and Syntactic Features (TS) The features in this group capture the surface-level identification of the usefulness and are listed as follows.

- *Text Statistics* – The aggregate statistics that can be extracted from the comments may also be good indicators for the usefulness of the comments. For instance, the longer comments are more likely to be useful because take more space to represent information and take longer time to be written. A higher number of nouns may indicate that the comment contains knowledge from different aspects. Based on similar assumptions like these, we use the aggregate statistics extracted from the text such as number of words (#WC), number of verbs, number of adverbs, and the average length of sentences (WPS). We collect statistics based on the POS tags to create a set of features such as percentages of verbs, adverbs, etc. We use the

Features	**Short Description**
Text Statistics and Syntactic Features (TS)	
Readability	measures how difficult the comment is to parse using the Gunning fog index [10]
Informativeness	measures the novelty of terms, t, of a comment, c, compared to other comments on the same object, calculated using:$\Sigma_{t\in c} tfidf(t, c)$
Punctuation Mark	counts the number of punctuation marks
Text Statistics	measures aggregate statistics extracted from the text #Words, #Verbs, #Adverb, WPS (average length of sentences)
Linkage Variety	counts the number of unique hyperlinks in a comment
Semantic and Topical Features (ST)	
Named Entities	counts the number of named entities that are mentioned in a comment
NE Types Variety	counts distinct types of named entities (such as person, place, date, etc.) that are mentioned in a comment
Topical Conformity	measures the distance between the topics of a comment and the topics belonging to other comments on the same object. We use the Jensen-Shannon (JS) divergence to measure the topic distribution distance of all comments on an object (A) compared to the comment's topic distribution (C). $D_{JS} = \frac{1}{2}(D_{KL}(C \parallel A) + (D_{KL}(A \parallel C)$ and KL divergence is calculated as: $D_{KL}(C \parallel A) = \Sigma C(i) log\frac{C(i)}{A(i)}$.
Sentiment Polarity	measures the sentiment/polarity of a comment as: $SenPolarity = \frac{PositiveScore+NegativeScore}{\#Words}$ We use LIWC for identifying positive and negative scores.
Subjectivity Tone	measures the subjectivity degree of a comment. We use Subjectivity Lexicon [25] to calculate subjectivity
User Topic Entropy	measures the topical focus of an author via the entropy of topic distributions of the author. We define entropy of topic distribution of all comments authored by an author, a_i as: $H(a_i) = -\Sigma_{j=1}^{n} p(t_{i,j}) \log p(t_{i,j})$, where t is a topic and n is #topics.
Psychological & Social characteristics of the content	identifies psychological dimensions: Leisure, Anger, Family, Friends, Humans, Anxiety, Sadness, Sexuality, Home, Religion, Relativity, Affective Process, and Self-reference scores [22]
User and Social Features (US)	
User Linkage Behavior	counts the number of unique hyperlinks posted by a user. A high linkage balance indicates that linkage is part of the commenting behavior of a user.
User Conversational Behavior	counts comments that contain a @reply
User Activity	measures different activities completed by a user: #Comments (counts the number of comments authored by the user), #UploadedObjects (counts the number of media objects uploaded by the user), #Favorite Objects (counts the number of media objects selected as favorite by the user)
User Social Relation	counts the number of contacts of the user and measures Prestige score (measures the number of the Flickr Commons members in the contact list of the user)

Table 2 Overview of Features

LingPipe toolkit[9] to obtain the relative POS taggers. We hypothesize that comments containing a higher number of words are likely to be useful [16, 9].

- *Linkage Variety* – The number of hyperlinks in a comment. The comments written by either experts or users with high relevant knowledge may tend to include the hyperlinks to external credible resources to support their text. Therefore, we hypothesize that the more links are contained in a comment, the more likely it is to be useful [6]. The example below shows a comment with high Linkage Variety, judged as useful by the coders:

  "There were 2 different Frances GALLWEY in Tramore. Here is the wife of William GALLWEY. 1901 census, 26 Circus, Bath, Somerset Phyllis DAVIES, Head, Widow, 88, Living on own means, born in Devon, Ugborough Frances K GALLWEY, Daughter, Married, 47, Living on own means, born in Yorkshire, Adlingfleet Jannette P GALLWEY, Granddaughter, Single, 17, Living on own means, born in Ireland plus 5 female servants, all born in Somerset.www.freebmd.org.uk Marriage, March quarter 1883, Bath William Joseph GALLWEY and Frances Kate T DAVIES thepeerage.com/p39134.htm Frances Kate Trelawner DAVIES was the daughter of Reverend Edward William Lewis DAVIES. She married William Joseph GALLWEY, son of Henry Gallwey and Maria Walsh, on 25 January 1883. She died on 29 March 1938. Ireland, Civil Registration Indexes, 1845-1958 Frances K T GALLWEY died in Waterford district, 1938."[10].
- *Informativeness* – This feature measures the novelty of terms used in the comment compared to other comments on the same object. Practically, we use the sum of the tf-idf, term frequency-inverse document frequency to calculate this feature:

  $$\Sigma_{t \in c} tfidf(t, c)$$

  Here, t is a term used in the comment denoted by c. The higher usage of novel terms in the comment may indicate that it brings more useful information. For that reason, we assume that comments with higher informativeness score are more informative and, therefore, they are likely to be useful [23].
- *Punctuation Mark* – The number of punctuation marks in the comment. Given that the emotion and a series of meaningless punctuations are frequently seen in comments that are not useful. Therefore, we assume that the number of punctuation marks may have impact on the usefulness of the comments.
- *Readability* – measures how difficult the comment is to parse by using the Gunning fog index [10]. We assume that comments with a higher readability score are likely to be useful, because they are easier to parse for humans. The example below shows a comment with high readability score, judged as useful by the coders:

[9] http://alias-i.com/lingpipe/

[10] Flickr photo - April 15, 1901 http://www.flickr.com/photos/nlireland/6933777014/comment72157629836757055

"After being the Boxing Champion of the World, Jimmy Clabby is said to have squandered over $500,000 in earnings, and was found dead of starvation in Calumet City during the Great Depression."[11].

Semantic and Topical Features (ST) Besides superficial identifications, we may get more insights of a comment by checking its semantics. The semantic information characterizing from different aspects may have various impact on the likelihood of a comment being useful regardless of its text structure. Furthermore, this group includes standard topical model features, which measure the topical concentration of the author of a comment and the topical distance of a comment compared to other comments made on the same object. Specifically, we analyze the following features:

– *Named Entities* – The number of named entities (NE) that are mentioned in a comment may give evidence on the usefulness. A comment with higher number of entities conveys more concepts that are known to the public. In practice, we use GATE toolkit[12] to ext extract NE related features in this group. We hypothesize that the more entities are identified, the more likely the comment is to be useful. The example below shows a comment with high number of name entities, which is judged as useful by the annotators: "[Claire L. Runkel (1890 – 1936) and Oscar F. Grab (18861958) were married on March 23, 1915, at the Ritz Carlton Hotel in New York City. Claire was the daughter of Herman Runkel (1853-1918) and Victoria Rebecca Runkel (nee Lopez) (1859-1927), of 150 W. 79th Street in New York City. Mr. Runkel was of the firm Runkel Brothers, chocolate manufacturers. Oscar F. Grab , born Oskar Grab, was an Austro-Hungarian immigrant, United States citizen, and fashion executive. He was a saloon passenger aboard Lusitania who saw the torpedo impact the ship on May 7, 1915. He saw lifeboats upset on the starboard side and jumped into the water instead of taking a chance in the lifeboats. He was rescued and survived the Lusitania disaster. His wife was not traveling with him. Oscar and Claire moved in with Claire's parents that October. The couple had two children, Victoria, born in 1916, and Donald born in 1923. Claire also authored a book, By 1928, Oscar's fashion company, O. F. Grab Company, was a million-dollar business that had branches in France and Belgium and was employing 250 people....]"[13].
– *NE Types Variety* – The number of distinct types of named entities (such as person, place, date, etc.) that are mentioned in a comment. More types of entities mentioned in a comment may indicate that the object is introduced from different aspects. Therefore, we hypothesize that a comment is more likely to be useful if the entities contained in it are more diverse in terms of their types. The previous example also shows the comments with high NE Types Variety.
– *Subjectivity Tone* – The fact or related background knowledge on an object tends to be described in an objective tone. So we assume the subjectiv-

[11] Flickr photo - Jimmy Clabby. Boxing `http://www.flickr.com/photos/library_of_congress/2163449292/comment72157603820313375`

[12] `http://gate.ac.uk`

[13] Flickr photo - (Clara Runkel) Mrs. Oscar F. Grab `http://www.flickr.com/photos/library_of_congress/6851810917/comment72157629260546153`

ity tone of a comment may impact the usefulness of the comment. By leveraging Subjectivity Lexicon [25], we can calculate the subjectivity of a comment. This enables us to construct the feature of *Subjectivity Tone* with the hypothesis, that a comment with objectivity tone is more likely to be useful. [9]. The example below shows a comment with high objectivity tone, which is judged as useful by the coders:

"Yes, this is the British pavilion by sir Edwin Lutyens."[14]

- *Psychological content characteristics* – We can extract psychological characteristics from the contents by using LIWC [22] for analyzing psychological characteristics. This can give us indicators in various dimensions, including leisure, anger, family, friends, humans, anxiety, sadness, sexuality, home, religion, relativity, affective process, and self-reference. The scores involving authors' mood, which may be represented by the scores of anger, sadness, may have impact on the usefulness of the comments. We can suspect that a comment with high sadness or anger scores might be written when the author was in a bad mood, therefore is likely to be biased. The example below shows a comment with high sad score, which is judged as non-useful by the coders:

  "Seeing alcohol being so wasted just makes me want to cry." [15]
- *Topical Conformity* – This feature measures the distance between the topics of a comment and the topics detected in other comments on the same object. An LDA model (Latent Dirichlet Allocation [5]), was trained to handle features that depend on topic models. To train the LDA model we aggregated all the comments on objects in our database into an artificial document to infer topic distribution and chose the following hyperparameters: $\alpha = 50/T$, $\beta = 0.01$ and T = 1,000[16]. Then, we used the Jensen-Shannon (JS) divergence to measure the topic distribution distance of all comments on an object A compared to the comment's topic distribution C.

$$D_{JS} = \frac{1}{2}(D_{KL}(C \parallel A) + (D_{KL}(A \parallel C)$$

  and KL divergence is calculated as:

$$D_{KL}(C \parallel A) = \Sigma C(i) log \frac{C(i)}{A(i)}$$

  The high topical conformity means the comment is closely related to the core message conveyed in the artificial document, and therefore is probably the characteristic of useful comments. For this reason, we hypothesize that the higher the topical conformity we find for a comment the more likely it is to be useful [24,23].

[14] Flickr photo - Paris Exposition: Hungarian Pavilion, Paris, France, 1900 `http://www.flickr.com/photos/brooklyn_museum/2486821878/comment72157613666119960`

[15] PROHIBITION DOCUMENTARY. `http://www.youtube.com/watch?v=OiYqFXmVAFg`

[16] We also experimented with a different number of topics (10, 100, and 500) for training the LDA model. However, our results — discussed in the Experiments section — have shown that training the LDA model using 1,000 topics is a most influential setting

- *User Topic Entropy* – The topical focus of an author measured by the entropy of topic distributions of a user may indicate whether she is focusing on some certain topics. This feature can be inferred via the whole set of comments she authored. To handle this feature, we again trained an LDA model [5]. For this purpose, we aggregated all the comments authored by each user in our database into one artificial user document to infer topic distribution by her and we chose the following hyper-parameters: $\alpha = 50/T$, $\beta = 0.01$ and T = 1,000[16]. Given the inferred distance topic distribution of each user, we define entropy of topic distribution of all comments authored by an author, a_i as:

$$H(a_i) = -\Sigma_{j=1}^{n} p(t_{i,j}) \log p(t_{i,j})$$

  Here, t is a topic and n is #topics. We assume the topical focus of users has influence on the usefulness of their comments.
- *Sentiment Polarity* – Previously, researchers found that the sentiment polarity has an impact on the usefulness of the comments [9, 6]. We construct this feature as following formula:

$$SenPolarity = \frac{PositiveScore + NegativeScore}{\#Words}$$

  The useful comments, which are informative, should be written with less emotion from the author. Therefore, we hypothesize the lower the sentiment polarity found in a comment, the more likely it is to be useful. The example below shows a comment with high Sentiment Polarity score, which is judged as non-useful by the coders:

  "Martins my namesake was great i will be great also. Hahahahaha!!! Interesting." [17]

User and Social Features (US) In addition to the before mentioned features, which describe characteristics based on syntactical and semantic information, we also look into the features that describe the context in which a comment was published. Due to limitations of access to this information, we apply a lightweight characterization of authors and their social contexts. We particularly analyze following features:

- *User Linkage Behavior* – The number of unique hyperlinks posted by a user. A high usage of linkage indicates that the author has the behavior of including hyperlinks. As mentioned above, using a hyperlink may support the comment. Here, we evaluate this usage by users. Therefore, we assume that the comments by users that use other resources as references are more likely to be useful.
- *User Conversational Behavior* – On social media platforms, users can interact with each other by writing a comment containing an @reply. The reply messages are frequently found to be questions to previous comments, simple answers to it, or even chat messages. Therefore, we assume that

[17] YouTube Video – Martin Luther King, Jr. - Mini Bio http://www.youtube.com/watch?v=3ank52Zi_S0

users that write comments to converse with other users are less likely to write useful comments.

- *User Activity* – We can measure the activities completed by a user from different aspects, e.g. the number of comments authored by the user, the number of media objects uploaded by the user, and the number of media objects marked as favorite by the user. The higher these indicators are the more active the user is on the platform. Inspired by [8,6], we construct these features and hypothesize that the more active the user is, the more likely the comments authored by her are seen as useful.
- *User Social Relation* – We measure the social relation of an author by two metrics: the number of contacts that she has and the Prestige score measured by the number of the influential contacts (such as Flickr Commons members) in the contact list of the user. We assume that users with a higher number of social interactions are more likely to write useful comments [19].

4 Data Acquisition

In this section we describe how we collect usefulness judgements for characterizing useful comments. We achieve this by building a dataset from real world comments harvested from Flickr Commons and YouTube, which provide free-text comments on media objects (video and photo) from a variety of people with different backgrounds and intentions, first by extracting those comments that have attracted a response by experts of cultural institutes, and second, by using a crowd-sourcing approach, setting up a user study, and requesting people to state if they consider that a certain set of comments could be useful for them. Finally in order to show how users' perception of usefulness is similar to experts' perception, we compare the characteristics of useful user-judged and expert-judged comments [3,4].

4.1 List of Topics

In order to analyze the correlation between usefulness and different topics of media objects, we first selected three types of topics: *event*, *person*, and *place*. We selected these three broader areas of topics because the identification of these topics is supported by a significant set of automated tools and approaches in a variety of application domains. Second, we used the history timeline of the 20th century provided by About.com to identify topics associated with the selected topics from each decade of the 20th century. The resulting topics included, among others, 'Irish civil war" and "1936 Olympics" as events,"old New York" and "old Edinburgh" as places, and "Neil Armstrong" and "Princess Diana" as people.

Photoset	Topic Type	Comments	Objects	Users
Library of Congress	Person, Event	27,603	9,029	4.343
Brooklyn Museum	Place	2,178	251	1,687
National Library of Ireland	Event, Person	1,740	135	470
New York Public Library	Place	251	98	151
National Gallery of Scotland	Place	257	32	201
NASA Collections	Person	103	28	82

Table 3 Summary statistics of dataset crawled from Flickr Commons [3]

4.2 Datasets

We built two datasets from real world comments harvested from Flickr Commons and YouTube:

- **Dataset1:** we crawled comments written on photos of six different cultural institutions on Flickr Commons. We searched Flickr Commons for photo-sets of each topic (when available) and selected photo-sets that have the highest number of comments on their photos. In one of the Library of Congress photo-sets (News in 1910), it is worth mentioning that many of the photos are of persons. Accordingly, photos which show only a photo of a person are separated by us from other photos which belonged to topics related to event, according to their titles. Also, in order to train a classifier and analyze users' features, we crawl all profile information of all users who wrote comments. Table 3 shows the summary statistics of the dataset.
- **Dataset2:** we compiled a dataset from real-world comments harvested from YouTube, searched YouTube for videos of each topic (when available), selected those with the highest number of views and comments (at least 100), and crawled 91,778 comments (the first $1,000$ for each topic) written for 310 different videos. For each comment we crawled all the available profile information for the author. Because of access restrictions to certain user profile fields and crawling limitations (e.g, max. 1000 comments) we couldn't build all the mentioned features in Section 2 for YouTube dataset (such as Informativeness or Topical Conformity and some user related features such as number of contact, prestige score, number of favorite objects, etc) in the feature engineering phase.
  In total for Flickr we crawled 33,273 comments written on 11,102 photos. For YouTube we crawled 91,778 comments (the first $1,000$ for each topic) written for 310 different videos. (Distribution of the comments across different topics is shown in Table 4.) As a result, we obtained comparable datasets from YouTube and Flickr for topics involving events, people, and places across different time periods starting in 1900.

Platform	Event	Place	Person	Total
Flickr	13,864	6,935	12,474	33,273
YouTube	50,654	6,908	34,216	91,778

Table 4 Summary statistics for datasets [4]

Platform	Total	Useful	Not Useful	Agree
Flickr	3,500	1,345 (38.42%)	2,155 (61.57%)	0.86
YouTube	5,000	414 (8.28%)	4,586 (91.72%)	0.72
ALL	**8,500**	**1,759 (20.69%)**	**6,741 (79.30%)**	**0.79**

Table 5 Manual coding results across platforms. Agreement scores are assessed based on Mean Fleiss' Kappa scores [4].

4.3 Collecting User Judgements for Defining Usefulness

We randomly selected 3,500 comments from Flickr and 5,000 from YouTube and coded them manually as being useful or non-useful. (As will be seen below, more comments were required from YouTube due to the low rate of useful comments.) See Table 5 for results of manual coding.

Coders were recruited via the CrowdFlower.com crowd-sourcing platform which distributed our task across different channels, such as Mechanical Turk or getPaid. Coders were asked to assist us to define useful comments, by showing each coder a comment and links to the related media object (Flickr photo or YouTube video). We asked coders to answer three objective questions, to ensure quality of coder responses. The answers to the first three questions can be computed automatically, and a fourth question addressed the usefulness of the comment. The first and second questions for both platforms were semantically the same but asked in two different ways. Inconsistency in answering the first two questions gives us the chance to exclude randomly selected answers. The first two questions for the Flickr user study are: 1-"how many Web links does the comment contain?", 2- "does the comment contain Web links"? The first two questions for the YouTube user study are: 1- "Is the length of the video short or long?" (more than two minutes is long, less than two minutes is short) 2- "how long is the length of the video?" The third question required writing a text-based answer, offering an additional chance to exclude data from non-serious coders. The main question (the fourth question) for the task was the following: *"Compared to the description provided by the uploader of the media object (located below the video or photo), is this comment useful for you to learn more about the content of the media object (video or photo)?"*. For each comment we collected three judgements.

In order to prepare a training-set for developing a usefulness classifier, first, we select $1,000$ user-judged useful comments with high agreements on being useful and $1,000$ comments with high agreements on being non-useful from our labeled data.

Second, we assess the mean values and standard deviations of each feature, as shown in Table 6. As expected, the average semantic and topical-based scores for comments which are judged as useful are different from those for

	Flickr				**YouTube**			
Features	**Mean-U**	**STD-U**	**Mean-N**	**STD-N**	**Mean-U**	**STD-U**	**Mean-N**	**STD-N**
Text Statistics and Syntactic Features (TL)								
Readability	06.05	04.07	05.70	03.54	09.12	07.87	05.46	05.38
#Punctuation Marks	77.76	131.4	77.10	214.7	25.02	28.63	32.06	44.17
#WC	41.70	49.41	09.32	12.52	41.17	31.77	19.82	19.79
#WPS	15.63	10.99	06.36	06.50	20.47	17.99	12.51	11.79
#Verb	09.06	08.61	09.05	11.38	13.61	06.99	14.34	11.02
#Adverb	02.91	04.81	05.10	10.30	04.54	05.52	04.80	07.37
Linkage Variety	01.72	01.82	0.521	05.92	–	–	–	–
Informativeness	14.50	21.91	05.02	06.37	–	–	–	–
Semantic and Topical Features (ST)								
#Name Entities	03.62	05.33	0.466	0.956	02.44	02.77	01.07	01.67
NE Types Variety	01.39	01.07	00.36	00.58	01.10	00.83	0.639	0.704
Topical Conformity	01.34	01.67	01.07	01.10	–	–	–	–
Sentiment Polarity	01.62	03.75	29.26	32.77	06.59	09.01	10.44	15.28
Subjectivity Tone	0.151	0.160	0.910	0.750	0.187	0.122	0.296	0.265
Sadness	0.190	0.880	0.160	0.940	0.411	01.29	0.562	04.09
Insight	0.150	01.56	0.096	0.810	01.48	02.35	01.66	03.90
Anger	0.369	01.74	0.197	01.80	01.91	05.92	02.41	07.29
Family	0.460	01.63	0.126	01.40	0.359	01.42	0.329	01.74
Friends	0.060	0.950	0.130	02.98	0.049	0.497	0.087	01.10
Humans	0.590	01.93	0.840	03.64	01.33	03.49	01.26	03.88
Health & Body	0.790	02.41	01.93	07.02	01.29	03.52	02.28	06.65
Sexual	0.065	1.086	0.970	05.10	0.356	0.528	01.06	05.00
Religion	0.409	02.86	0.103	01.21	0.404	0.30	0.61	03.50
Leisure	01.30	02.99	0.460	02.51	01.29	02.75	01.57	05.60
Swear	0.058	0.087	0.198	0.682	0.216	01.44	01.33	06.12
Home	0.450	01.74	0.180	01.35	0.091	0.515	0.167	01.07
Relativity	12.86	09.18	06.14	09.87	12.61	08.46	10.23	11.07
Certainty	0.616	1.980	1.290	6.750	01.54	02.81	01.97	05.37
Tentative	01.79	03.65	01.21	03.98	02.07	03.22	02.00	04.72
Self-reference	01.02	2.587	02.27	05.42	01.24	02.73	03.08	06.19
User Topic Entropy	04.74	01.67	04.34	02.69	–	–	–	–
User and Social Features (US)								
User Linkage Behavior	758.0	1225	09.93	88.44	–	–	–	–
User Conversational Behavior	0.480	02.35	19.20	33.65	0.522	0.501	0.392	0.488
#UploadedObject	20250	3869	1390	3134	11.17	64.94	05.46	34.48
#FavoriteObject	243.5	220.5	269.1	219.5	–	–	–	–
#Contact	179.1	261.7	204.6	283.6	–	–	–	–
Prestige score	04.96	09.61	01.62	4.274	–	–	–	–

Table 6 The comparison of the mean and standard deviation values of each feature between useful (U) and non-useful (N) comments. The underlined values point out considerable differences between useful (U) and non-useful (N) comments [3].

non-useful comments. The *Sentiment Polarity* and *Subjectivity Tone* scores for comments which are judged as non-useful are much higher than those for useful comments. Comparing NE-dependent semantic features reveals that useful comments generally contain more entities (2-3 entities) than non-useful comments (0-1 entity). The *NE Type Variety* (only person, organization, location, and date are considered) is higher for the useful comments than for the non-useful comments. Among the psychological characteristics of the content, those which are judged as useful such as the average *Insight*, *Friends*, *Health & Body*, *Religion*, *Swear* and *Sexual* scores for comments, which are judged as useful, are different from those for non-useful comments. With regard to user and social features, for Flickr the user *Linkage Behavior* and *Prestige* scores for comments, which are judged as non-useful are much higher than for those for useful comments. For YouTube the number of *UploadedObject* by a user is potentially a good indicator. For features related to the text statistics and syntactic we observe that regardless of whether the comments are useful or not, the ratios of comments with higher text statistic scores are almost the same. For example, it seems that the presence of punctuation marks is not necessarily an indicator of usefulness. However, the presence of hyperlinks (*Linkage* score) and the number of words per sentence (WPS) are potentially good indicators.

4.4 Collecting Expert Judgements for Defining Usefulness

With regard to comments written on photos of the Library of Congress (LOC), we notice some of these comments are commented upon by the LOC experts[18]. In order to ensure that these comments are useful for LOC, we ask LOC staff members why they comment back. They confirm that commenting back is one indicator of a useful comment: "all Flickr comments are being read by LOC staff. The vast majority of comments is useful, but we only have the resources to comment back when we verify that a suggested change was on target, so that the Flickr users know that their information is making a difference.".

Based on these observations, first we crawl all comments written by LOC staff and containing terms such as "thanks", "thank you", etc. Second, in order to find related comments to these comments, we use the crowd-sourcing approach and we ask coders to assist us in defining relevant comments. We use CrowdFlower.com which is a crowd-sourcing platform, showing each coder a comment written by LOC staff and links to the related Flickr photo and asking them to find all relevant comments to LOC experts' comments. In total we gather comments amounting to 2,068, which we presume to be considered useful by experts. It is worth mentioning that LOC experts have not explicitly classified comments as useful and non-useful. This means that comments which in our study are inferred as "non-useful" might be useful for other contexts and the term "useful" is a term that we use in our study. Furthermore, in order to compare characteristics of user-judged with expert-judged useful comments we randomly selected 1000 expert-judged useful comments and we selected $1,000$ user-judged useful comments with high agreements on being useful.

Second, we assess the mean values and standard deviations of each feature for expert-judged comments. Table 7 shows in detail these values in comparison with user-judged useful comments. This table shows the mean and standard deviations of almost all features from both datasets are in the same range. This result suggests that the characteristics of user-judged comments are very similar to characteristics of expert-judged useful comments and therefore the

[18] These are user accounts which have the pattern "Name (LOC P&P)" and use the Library of Congress logo

Features	**Mean-U**	**STD-U**	**Mean-E**	**STD-E**
Text Statistics and Syntactic Features (TL)				
Informativeness	14.50	21.91	15.36	25.47
Readability	06.05	04.07	06.78	04.31
#Punctuation Marks	77.76	131.4	185.9	219.0
#WC	41.70	49.41	48.60	62.59
#WPS	15.63	10.99	17.53	12.82
#Verb	09.06	08.61	07.60	07.47
#Adverb	02.91	04.81	01.59	03.08
Linkage Variety	01.72	01.82	03.87	03.76
Semantic and Topical Features (ST)				
#Name Entities	03.62	05.33	06.93	08.50
NE Types Variety	01.39	01.07	01.83	01.01
Topical Conformity	01.34	01.67	01.56	01.19
Sentiment Polarity	01.62	03.75	01.78	03.49
Subjectivity Tone	0.151	0.160	0.105	0.078
Sadness	0.190	0.880	0.143	0.659
Insight	0.150	01.56	0.965	02.33
Anger	0.369	01.74	0.336	01.09
Family	0.460	01.63	0.538	01.64
Friends	0.060	0.950	0.055	0.541
Humans	0.590	01.93	0.596	01.74
Health & Body	0.790	02.41	0.234	01.14
Sexual	0.065	1.086	0.035	0.310
Religion	0.409	02.86	0.303	01.56
Leisure	01.30	02.99	01.18	02.84
Swear	0.058	0.087	0.014	0.272
Home	0.450	01.74	0.225	0.923
Relativity	12.86	09.18	11.61	09.58
Certainty	0.616	1.980	0.425	2.217
Tentative	01.79	03.65	01.13	02.58
Self-reference	01.02	2.587	00.61	1.931
User Topic Entropy	04.74	01.67	04.75	01.34
User and Social Features (US)				
User Linkage Behavior	758.0	1225	771.0	1378
User Conversational Behavior	0.480	02.35	0.520	02.35
#UploadedObject	20250	3869	30250	7869
#FavoriteObject	243.5	220.5	298.4	247.5
#Contact	179.1	261.7	184.0	192.0
Prestige score	04.96	09.61	04.74	09.64

Table 7 The comparison of the mean and standard deviation values of each feature between user-judged (U) and expert-judged (E) useful comments [3].

non-useful comments (labeled in our study) can be assumed to be non-useful from both perspectives.

5 Experiments

In this section, we introduce the process of building the learning-based "usefulness" classifier and evaluate it on the manually coded comments. Given the comments on which the usefulness is estimated, we calculate all the features introduced in Section 3 and attempt to build the classifier. The classifier can then be automatically used as an inference method to predict whether a comment is useful or not. We report on the estimation performance by applying different machine learning algorithms. Next, we evaluate the importance of the features that can be interpreted from the coefficients of the classifier. Finally,

Features	Classifier	Flickr				YouTube			
		P	**R**	**F1**	**ROC**	**P**	**R**	**F1**	**ROC**
TS	LR	0.76	0.75	0.75	0.85	0.56	0.56	0.56	0.60
	NB	0.74	0.71	0.71	0.77	0.60	0.59	0.59	0.65
ST	LR	0.84	0.85	0.84	0.93	0.66	0.72	0.68	0.71
	NB	0.81	0.80	0.79	0.89	0.62	0.87	0.71	0.72
US	LR	0.79	0.60	0.68	0.80	0.58	0.54	0.56	0.53
	NB	0.71	0.66	0.65	0.80	0.64	0.53	0.53	0.44
TS + ST	LR	0.85	0.85	0.85	0.89	0.68	0.72	0.70	0.72
	NB	0.79	0.79	0.79	0.88	0.63	0.84	0.72	0.72
ST+ US	LR	0.85	0.85	0.85	0.93	0.67	0.66	0.67	0.71
	NB	0.84	0.83	0.83	0.92	0.61	0.81	0.70	0.69
TS+ US	LR	0.84	0.83	0.83	0.90	0.62	0.67	0.64	0.67
	NB	0.80	0.77	0.77	0.86	0.61	0.87	0.71	0.72
ALL	**LR**	0.87	0.90	**0.89**	**0.94**	0.66	0.74	0.70	0.72
	NB	0.84	0.83	0.83	0.91	0.65	0.83	**0.73**	**0.72**
Baseline1	LR	0.61	0.53	0.57	0.59	0.51	0.50	0.50	0.52
Baseline2	LR	0.65	0.80	0.72	0.77	0.55	0.70	0.61	0.59

Table 8 Results from the evaluation of classification algorithms with different feature settings (**bold** indicates the top F1 and ROC scores for each dataset) [3,4].

we provide an interpretation of to what extent the commenting culture of a platform influences the performance of the usefulness classifier.

5.1 Usefulness Classifier

Experimental Setup For training the usefulness classifier, we selected a balanced set of 1,000 USEFUL comments and 1,000 NOT USEFUL comments from the Flickr data; we selected 400 of each class from the YouTube data. Each of these comments has been judged at least three times, by different coders. Moreover, to ensure the quality of the judgements, the comments may be selected only if at least two out of three coders had an agreement. Practically, we have employed two machine learning algorithms, logistic regression (LR) and Naive Bayes (NB), to build the classifiers. Classifiers were trained by using different combinations of the feature groups described in Section 3. For evaluation, we focus on four measures: precision (P), recall (R), F1-measure (F1), and area under the Receiver Operator Curve (ROC). Besides the proposed usefulness classifiers, we designed two baseline approaches for comparison purposes:

Baseline 1 predicts usefulness by using the feature of INFORMATIVENESS. This feature is demonstrated by Wagner et al. [23] to be an influential feature for predicting the attention level of a posting in online forums [4].

Baseline 2 predicts usefulness by using the feature of SUBJECTIVITY TONE, which is a particularly strong baseline as a result of our feature analysis study [4].

Results of Evaluations of Different Classifiers Table 8 provides an overview of both the estimation performances of the two baselines and classifiers trained with different combinations of the feature groups by using two machine learning algorithms. The results demonstrate the effectiveness of using semantic (ST) and user-related (US) features for inferring useful comments.

In particular, for both Flickr and YouTube dataset, the classifiers created by using author and semantic features outperform the models trained with text features (TS) by using the algorithm of either Logistic Regression or Naive Bayes. Specifically for Flickr dataset, we are able to achieve an F1 score of 0.89, coupled with high precision and recall, when using the Logistic Regression classifier in combination with all features. However, we find a lower level of F1 score (0.70) when using the same machine learning algorithm on the YouTube dataset. On the contrary, we are able to achieve an F1 score of 0.73 by applying the algorithm of Naive Bayes. ROC measures show similar levels of performance for the algorithms of both Logistic Regression and Naive Bayes over the two datasets.

In general, we found the performance on YouTube dataset is lower than on Flickr dataset due to the fact that we also did not have high agreement among coders in manual coding. Another reason may be that we have not constructed all the author-related features (US) due to the API limitation.

5.2 Influence of Features on Usefulness Classifier

Experimental Setup Having analyzed the influence of using different combinations of features groups on the estimation performance, we now evaluate the importance of individual features for inferring the usefulness of comments for both datasets.

To investigate how the features were associated with the usefulness of comments, we examine the coefficients of the best-performing Logistic Regression model (using *ALL* groups of features). Table 9 lists the coefficients of 20 features that are highly ranked in terms of Information Gain Ratio (IGR). The features with positive coefficients are positively correlated to the usefulness while the negative coefficients are negatively correlated to the usefulness. Following, we analyze these results and try to validate our hypotheses made in Section 3.

Results of Influential Features The top-ranked features from two datasets are both dominated by Semantic and Topical (ST) features. More specifically, coefficient ranks show that comments that express emotional and affective processes of the author (higher *Subjectivity Tone*, *Sentiment Polarity*, *Anger*, *Sadness*, *Swear*, and *Anxiety* scores) are more likely to be inferred as NOT USEFUL. *Subjectivity Tone* is a very good indicator for both platforms. Higher *Subjectivity Tone* has negative impact on the usefulness classifier. Therefore, we have the hypothesis made in Section 3 validated. Furthermore, comments with offensive language (higher *Swear* score) are more likely to be inferred as NOT USEFUL. An analysis of the *Swear* and *Anger* scores between different platforms shows that YouTube contains more offensive language. Therefore, the *Swear* and *Anger* scores for YouTube are more negative than the Flickr swear score. This can be explained by that more frequent emotional comments are posted on YouTube, while on Flickr this is not the case. Besides, the ranks show that comments that have higher *#Named Entities*, *NE Type Variety*,

Rank	Flickr		YouTube	
	Feature	**Coefficient**	**Feature**	**Coefficient**
1	**ST-Subjectivity Tone**	**-3.828**	**ST-Subjectivity Tone**	**-1.499**
2	**ST-Sentiment Polarity**	**-1.157**	**ST-#Name Entities**	**0.157**
3	**ST-NE Types Variety**	**0.550**	**ST-Self-reference**	**-0.126**
4	**US-User Linkage Behavior**	**0.025**	**ST-Swear**	**-0.167**
5	**ST-#Name Entities**	**0.211**	**ST-Sentiment Polarity**	**-0.014**
6	**ST-Self-reference**	**-0.148**	**ST-NE Types Variety**	**0.042**
7	**ST-User Topic Entropy**	**-0.049**	**ST-Anger**	**0.055**
8	**ST-Insight**	**0.049**	**ST-Tentative**	**0.051**
9	**ST-Swear**	**-0.045**	**US-#UploadedObject**	**0.084**
10	**TS-Linkage**	**0.173**	**TS-Future Verb**	**-0.143**
11	**US-User Conversational**	**-0.023**	**ST-Certainty**	**-0.012**
12	**ST-Certainty**	**-0.032**	**US-Author Conversational**	**0.027**
13	**TS-Future Verb**	**-0.043**	**ST-Anxiety**	**-0.134**
14	**TS-Impersonal-pronoun**	**0.025**	**TS-Impersonal-pronoun**	**-0.013**
15	**US-Prestige score**	**0.060**	**ST-Friend**	**-0.032**
16	**ST-Religion**	**0.089**	**ST-Religion**	**0.016**
17	**ST-Sadness**	**-0.075**	**ST-Sadness**	**0.036**
18	**ST-Sexual**	**-0.014**	**ST-Sexual**	**-0.059**
19	**ST-Family**	**0.016**	**ST-Home**	**-0.355**
20	**ST-Relativity**	**-0.006**	**ST-Family**	**-0.019**

Table 9 Top-20 features for each platform and related coefficient ranks derived from the Logistic Regression model. Features are ranked based on Information Gain Ratio [3,4].

and *Linkage* scores contain potentially interesting information and are likely to be inferred as useful. Therefore, we confirmed the assumption made for Named Entity related features.

We have constructed a series of features with the name of Psychological content characteristics (see Section 3) by using LIWC. The usage of terms in LIWC's `insight` category (such as think, know, consider) shows positive correlation with usefulness on Flickr dataset. This is in line with the relatively high difference of this feature between useful and not useful comments. Furthermore, terms in LIWC's `certainty` category (such as always, never) have a negative impact on the model. This might be due to the fact that authors who are assertive and express certainty tend to be seen as more subjective and less analytical. In contrast, using terms in LIWC's `tentative` category (such as maybe, perhaps, guess) shows that authors make less claims as to the correctness or certainty of their comments and such comments are likely to be determined useful.

It is interesting to note that *Readability* features are assigned little weight by the classifier. We suspect that this is because, while comments that are longer and contain more complex words are less "readable" based on the Gunning fog score, such comments are not necessarily less useful than comparatively shorter or less complex comments. Therefore, our hypothesis for the feature of "Readability" is not supported by the result.

With regard to User & Social (US) features, *User Linkage Behavior* is a good indicator showing that authors may diligently cite references for the information they provide. This increases reliability when inferring such comments as useful. Similarly, we note that a higher *Linkage* score has a positive impact on the usefulness inference, which is in line with the correlation of User Linkage Behavior score. Consequently, we can confirm the hypotheses made for these two features. A higher score of *Self-reference* and a higher *User Conversational* score have a negative impact. This suggests that authors who mostly use systems to converse and describe their personal experiences do not write useful comments. Again, we have validated our thoughts while constructing these two features. Interestingly, a higher *User Topical Entropy* score of authors has a negative impact on the usefulness inference. This indicates that authors with a higher entropy have a lower topical focus and therefore write a comment with a lower level of focus and knowledge about the specific topic. Therefore, their comments are likely to be inferred as NOT USEFUL.

Results of Iteratively Appending Features In order to observe the impact of iteratively appending features on classification performance, we conduct a further experiment to investigate how the performance of the classifiers changes as the top-ranked features are increasingly added for training. In particular, we apply the Logistic Regression algorithm for training - based on its optimum performance during the model selection phase - and trained the classifier using the training split from the first dataset. In Figure 1 we can see how the performance of the classifier changes with more and more top ranked features. The result shows the classifier can achieve about 70% and 80% of best performance in terms of F1 and ROC respectively with only one feature. With top 7 features, the trained classifier can already achieve about 90% and 95% of the optimal F1 and ROC respectively. By further adding features ranked lower, we observe similar levels of performance.

The results of this analysis show that a few relatively straightforward features can be used to characterize and infer the usefulness of comments. It is interesting to note that many text features, while being positively aligned with usefulness inference, do not belong to the most important features. On the contrary, Semantic and Topical features (ST) play important roles.

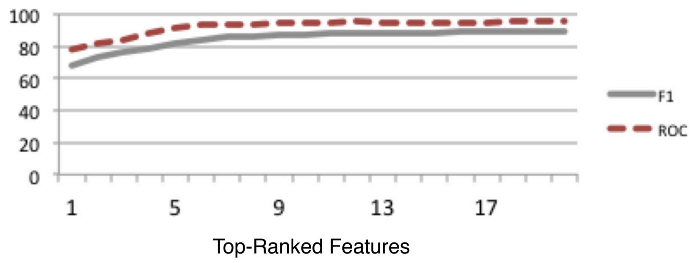

Fig. 1 Performance results of classification using top-20 features (Results of Iteratively Appending Features) [3].

		Person		Place		Event	
Platform		**All**	**Person**	**All**	**Place**	**All**	**Event**
Flickr	**F1**	0.82	0.89 *	0.73	0.87 *	0.93	0.94
	ROC	0.93	0.97	0.93	0.97	0.96	0.96
YouTube	**F1**	0.70	0.80 *	0.67	0.74	0.82	0.84 *
	ROC	0.74	0.89	0.75	0.83	0.85	0.88

Table 10 Results from the evaluation of usefulness classifiers for different object types. *All* is the type-neutral classifier which is trained on data corresponding to all topic types of objects. For each topic, we also show the performance of type-specific which is trained on data corresponding to all topic types of objects. "*" indicates a significant difference (p < 0.01) [4].

5.3 Influence of Topic on Classification

Experimental Setup In all results reported so far, we have largely ignored the particular characteristics of the objects commented upon. To explore how the importance of features varies for objects of different topics being commented upon, we divide the dataset into three splits according to the object topic types, Person, Place, and Event.

For each type of topic, we then compare the performance of two classifiers: a *type-specific classifier*, which we train by using only data of the same type as the test set, and a *type-neutral classifier*, which we train by using the whole dataset. The result indicates whether it makes sense to build the classifier for a certain type of topic of object.

Results from the evaluation of usefulness classifiers for different topics The performance results for type-specific and type-neutral classifiers are given in Table 10. We find that, in general, performance is better when the classifier is trained on comments of a single type, i.e., the classifier is type-specific, whereas performance is worse when the type is ignored, i.e., the classifier is type-neutral. We additionally perform three Pearson's Chi-squared tests between the prediction results of each classifier for each topic. In Table 10, "*" indicates a significant difference at a $p < 0.01$ level for some types. We can conclude that it at least makes sense to build a specific model for the object type of Person or Place.

Furthermore, we investigate the importance of features for each topic type of object with regard to usefulness inference. Table 11 shows detailed coefficient ranks for different models of three types of topics. Our discussion of the results focuses on the difference between the classifiers derived for each of the topic types. An analysis of the most important features among different type of objects (Person, Place, and Event) shows some differences. The major differences appear among the *Psychological characteristics of the content*, but a few differences appear among other semantic and user features. There is no significant difference among text features.

More precisely, coefficient ranks show that comments related to the type of topic, Person and Event express the author's emotional and affective processes more. These contribute to a comment being classified as NOT USEFUL. An analysis of the *Subjectivity Tone* among different topics shows that the

Subjectivity Tone for objects related to Person is higher than for other types. This can be explained by that authors of NOT USEFUL comments tend to use a subjective tone. An analysis of the *Swear* score among different topic types shows that the *Swear* score for the topic type, Person is the most negative one. With regard to the objects related to Event, the *Swear* score is more negative than for topics related to place.

For objects related to Person, the scores of *Family*, *Health* and *Body* implies that these features have a positive impact on the usefulness of the comments. This might be due to the fact that people describe more about various health and physical aspects of a person on these objects within the contributions that are considered to be useful. Furthermore, they describe the background of family members of the target person. This information may be useful information for others.

It is interesting to note that, for the objects related to Place, *Relativity* scores have a positive impact on the usefulness of the comments. However, *Friend* and *Family* scores have a negative impact. This might be due to the fact that the description of various physical phenomena and motion processes on the topic type, Place is actually not contributing to the explanation of the features but simply appears rather for other purposes. Therefore, giving information about friends and family for an object with topic related to Place is NOT USEFUL.

With regard to objects with the type of Event, we found the classifier is the most similar to *type-neutral classifiers*. The reason behind this is probably that the comments often includes information about both topic types related to Person and Place. This means that a object related to Event is often also related to Person, Place or both. Therefore, the coefficient ranks are influenced by the two other topics. For example, the *Relativity* score that includes physical place and motion has a positive impact in the type-specific model for topic types related to Place and Event, while it has a negative impact for the model for topic type related to Person.

5.4 Influence of Commenting Culture of Platforms on Characteristics of Useful Comments

As shown in Table 8, the result demonstrates that different platforms (Flickr and YouTube) lead to performance differences in usefulness classification. For all topic types (Place, Person, and Event), the performance of usefulness classifiers derived from Flickr platform is higher than that from the YouTube platform. Besides, the data limitation mentioned before (see Section 5.1), this may also be caused by cultural differences in commenting behaviors.

Furthermore, we investigate each feature by comparing the difference in coefficients in usefulness classifiers built with two platforms. For Flickr, we note a higher *Contact* score does not have a negative impact. However, a *Prestige* score has a positive impact. This indicates that having influential contacts in the contact list is more important than having a higher number of contacts.

Flickr				**YouTube**			
Feature	**Place**	**Person**	**Event**	**Feature**	**Place**	**Person**	**Event**
ST-Subjectivity Tone	-4.271	-6.228	-3.406	**ST-Subjectivity Tone**	-0.129	-2.386	-2.002
ST-Sentiment Polarity	-0.157	-0.223	-0.647	**ST-#Name Entities**	0.049	0.124	0.209
ST-NE Types Variety	-0.138	0.113	0.776	**ST-Self-reference**	-0.148	-0.46	-0.360
US-User Linkage Behavior	0.046	0.003	0.002	**ST-Swear**	-0.002	-0.571	-0.145
ST-#Name Entities	0.203	0.109	0.201	**ST-Sentiment Polarity**	-0.023	-59.734	-0.173
ST-Self-reference	-0.161	-0.136	-0.177	**ST-NE Types Variety**	-0.109	-0.175	0.328
ST-User Topic Entropy	-0.112	-0.302	-0.059	**ST-Anger**	-0.188	-0.138	-0.131
ST-Insight	-0.124	0.081	0.064	**ST-Tentative**	0.171	0.051	0.120
ST-Swear	-0.005	-90.427	-3.363	**US-#UploadedObject**	0.015	1.556	0.014
TS-Linkage	0.084	3.028	0.610	**TL-Future Verb**	-0.426	-0.182	-0.298
AS-User Conversational	-0.086	-0.086	-0.066	**ST-Certainty**	0.023	-0.034	-0.003
ST-Certainty	0.110	0.042	-0.054	**US-User Conversational**	-0.154	-0.484	0.083
TS-Future Verb	-0.071	-0.027	-0.027	**ST-Anxiety**	-0.216	-0.339	0.008
TS-Impersonal-pronoun	-0.052	-0.040	-0.042	**TS-Impersonal-pronoun**	-0.018	0.041	-0.087
US-Prestige score	0.162	0.005	0.070	**ST-Friend**	-0.519	-0.046	-0.011
ST-Religion	0.361	0.322	0.089	**ST-Religion**	0.046	-0.017	0.021
ST-Sadness	-0.110	-0.403	-0.038	**ST-Sadness**	0.325	-0.218	0.289
ST-Sexual	-1.306	-0.812	-0.284	**ST-Sexual**	-0.007	-0.175	-0.059
ST-Family	-0.196	1.111	-0.004	**ST-Home**	-1.760	0.692	-0.611
ST-Relativity	0.163	-0.160	0.029	**ST-Family**	-0.233	0.352	0.031

Table 11 Top-20 features for each platform and related coefficient ranks derived from the Logistic Regression model. Features are ranked based on Information Gain Ratio [3,4].

For YouTube, users with a higher number of uploaded objects are more likely to write useful comments. This does not apply to Flickr. No comparison can be made between YouTube and Flickr on contact related features due to the lack of crawled data from YouTube

The above experimental results indicate that:

1. There are a few relatively straightforward features that can be used to infer usefulness of user-generated comments;
2. An analysis of the important features across different platforms and different object types reveals that when inferring usefulness, the impact of features varies slightly;
3. The major differences appear among the psychological and social features (derived from LIWC) of the content. Therefore, a classification model should be trained that takes the topic of media object into account for building type-specific usefulness classifiers with higher accuracy;
4. The commenting cultures on different social media platforms are different. Therefore, a classification model should be trained that takes the commenting culture of a platform into account for building the usefulness classifiers.

6 Discussion

We have conducted an analysis of user-generated comments on media objects of different social media platforms to examine the characteristics of useful comments and identify the important key features of comments for inferring usefulness. In order to achieve these goals, we have analyzed three different sets of features: text statistics and syntactic, semantic and topical, and user and social.

Our experimental findings show that Semantic and Topical features play important roles for inferring the usefulness of comments. For characterizing and inferring the usefulness of comments, a few relatively straightforward features can also be used. Comments are more likely to be inferred as useful when they contain a higher number of references, a higher number of Name Entities, a lower self-reference and affective process (lower sentiment polarity, lower subjectivity tone, swear score, etc). Therefore, we suggest that a commenting system should urge users to define references [14], adding unambiguous users-verified concept references to social media comments. This in turn has a positive impact on the usefulness of comments.

An analysis of the users' features shows the likelihood for inferring the usefulness of a comment may be increased by leveraging users' previous activities. Therefore, we believe that by designing a commenting service, designers should take this fact into account when designing users' profile pages. This also implies that useful comments do not result when users mostly comment to converse and to describe their personal experiences (higher self-reference score). Furthermore, an analysis of the usage of different terms indicates that insightful and tentative terms indicate a positive correlation with usefulness, while certainty terms do not.

An analysis of the important features among different topics (place, person, and event) indicates that when inferring the usefulness of comments, the influence of features varies slightly according to the topic areas of media objects. More emotion may be expressed and more offensive language may be used when writing comments about topics related to persons and events. Such comments are more likely to be inferred as non-useful. When writing about topics related to person, users describe more about the background of family members, their health, and physical characteristics of the person. This information may be useful information for other people. Similarly, writing about topics related to place when more physical phenomena and motion processes are described may be seen as useful information by other users. On the contrary, information about family tends to be considered non-useful by other users. Therefore, being able to determine the topic area of a media object prior to inferring usefulness helps to classify useful comments with higher accuracy.

Furthermore, our results demonstrate that different platforms (Flickr and YouTube) lead to different usefulness classification results and the influence of features may vary according to the commenting cultures of platforms. Therefore, for a more accurate classification of useful comments, a classification model should be trained with regard to the commenting culture of a platform.

6.1 Limitation and Future Work

Considering the results of our experiments and analyzing the state-of-the art, we discover a number of limitations, and addressing these creates natural avenues for future work.

1. *Biases of judgements by crowd:* The wisdom-of-the-crowd approach simply allows all users to vote on (thumbs up or down, stars, etc.) or rate others' content. However, this approach avoids an explicit definition of usefulness. Crowd-based voting is influenced by a number of biases such as a "rich get richer" phenomenon that may distort accuracy [17].
2. *Removal of control from end-users:* This work introduces a machine-based approach which uses a trained classifier to rank comments based on a set of majority-agreed labeled comments. Some of the biases that arise due to voting are avoided by this approach. Using this approach, however, removes control from the end-users. As a result, individual viewers do not have the opportunity to personalize the ranking based on their preferences.
3. *Various annotating cultures in different platforms:* Our results related to usefulness identification experiments demonstrate that different platforms (Flickr and YouTube) lead to different usefulness classification results and the influence of features may vary according to the commenting cultures of platforms. Therefore, training a classifier cannot be appropriate for different platforms with different commenting cultures.
4. *Complexity of usefulness:* Automatic ranking of comments by "usefulness" is generally complex, mainly due to the subjective nature of "useful". In addition, even human raters find it difficult to agree on the usefulness of comments [4]. Moreover, usefulness for an individual confounds and blends together two aspects: "relevancy" of the comment to what the user has in mind or the information she is looking for, and "personal interest" in the comment, thus attracting her attention. These should be treated separately. For example, a user who intends to look for emotional content may look for comments where the content is relevant to affectivity. However, this does not necessarily mean that this user has any personal interest in particular comments which are relevant to affectivity. As a result, it is important that systems take into consideration both these dimensions of usefulness and help individuals adapt ranking based on the particular objective which users happen to have in mind.
5. *Comments as short texts:* Many available approaches propose strategies for extracting topics by enriching the semantics of individual posts [28] and enabling users to explore topics in order to filter content with regard to their interests. However, comments are often very brief and topics discussed alongside comments are very noisy. Furthermore, as comments have multiple explicit dimensions (such as language tone, physiological aspects, etc), grouping them exclusively based on topic results in a single imperfect faceted ranking does not enable users to rank comments with regard to other potentially useful facets. Therefore, a system which combines higher level features alongside topic classification is desirable.

In the future, we will explore a technique for optimizing the ranking of comments by providing adaptive faceted ranking of comments. In this way, end-users can identify comments of interest and focus on their experiences with regard to both dimensions of usefulness (relevancy and interestingness).

7 Acknowledgments

This work was supported in part by a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Program (PIOF-GA-2009-252206). We also thank members of the Library of Congress and especially Helena Zinkham for their insightful comments and advice on the result of the LOC project on Flickr Commons and the anonymous reviewers for feedback and suggestions.

References

1. E. Agichtein, C. Castillo, D. Donato, A. Gionis, G. Mishne, E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media with an application to community-based question answering. In *Proceedings of WSDM*, 2008.
2. H. Becker, D. Iter, M. Naaman, and L. Gravano. Identifying content for planned events across social media sites. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12. ACM, 2012.
3. E. Momeni, K. Tao, B. Haslhofer, and G. Houben. Identification of Useful User Comments in Social Media: A Case Study on Flickr Commons. In *Proceedings of the 13th ACM/IEEE Joint Conference on Digital Libraries*, JCDL '13. ACM, 2013.
4. E. Momeni, C. Cardie, and M. Ott. Properties, Prediction, and Prevalence of Useful User-generated Comments for Descriptive Annotation of Social Media Objects. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, ICWSM '13. AAAI, 2013.
5. D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *the Journal of machine Learning res*, 2003.
6. C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *the 20th international conference*, WWW, 2011.
7. C. Danescu-Niculescu-Mizil; G. Kossinets; J. Kleinberg; and L. Lee 2009. How opinions are received by online communities: a case study on amazon.com helpfulness votes. In *Proceedings of the 18th international conference on World wide web*, WWW '09.
8. N. Diakopoulos, M. De Choudhury, and M. Naaman. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, CHI '12. ACM, 2012.
9. A. Ghose and P. G. Ipeirotis. Designing novel review ranking systems: predicting the usefulness and impact of reviews. In *ICEC '07: Proceedings of the ninth international conference on Electronic commerce*, 2007.
10. R. Gunning. *The Technique of Clear Writing.* McGraw-Hill, New York, 1952.
11. C. E. Hall and M. A. Zarro. What do you call it?: a comparison of library-created and user-created tags. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, JCDL '11. ACM, 2011.
12. H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, 2007.
13. F. M. Harper, D. Moy, and J. A. Konstan. Facts or friends?: distinguishing informational and conversational questions in social q&#38;a sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 2009.
14. B. Haslhofer, W. Robitza, C. Lagoze, and F. Guimbretiere. Semantic tagging on historical maps. In *ACM Web Science 2013*, Paris, France, May 2013. ACM.
15. Y. Kammerer, R. Nairn, P. Pirolli, and E. H. Chi. Signpost from the masses: learning effects in an exploratory social tag search browser. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 625–634, New York, NY, USA, 2009. ACM.

16. S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, 2006.
17. J. Liu, Y. Cao, C. Y. Lin, Y. Huang, and M. Zhou. Low-Quality Product Review Detection in Opinion Summarization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.
18. Y. Liu, J. Bian, and E. Agichtein. Predicting information seeker satisfaction in community question answering. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2008.
19. Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi. Exploiting social context for review quality prediction. In *Proceedings of the 19th international conference on World wide web*, WWW '10, 2010.
20. K. Seki, H. Qin, and K. Uehara. Impact and prospect of social bookmarks for bibliographic information retrieval. In *Proceedings of the 10th annual joint conference on Digital libraries*, JCDL '10, 2010.
21. B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08. ACM, 2008.
22. Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. 2010.
23. C. Wagner, M. Rowe, M. Strohmaier, and H. Alani. What catches your attention? an empirical study of attention patterns in community forums. In *ICWSM*, 2012.
24. K. Q. Weinberger, M. Slaney, and R. Van Zwol. Resolving tag ambiguity. In *Proceedings of the 16th ACM international conference on Multimedia*, MM '08. ACM, 2008.
25. T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
26. S. Siersdorfer, ; S. Chelaru; W. Nejdl; and J. San Pedro 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.
27. C.-F. Hsu; E. Khabiri; and J. Caverlee 2009. Ranking comments on the social web. In *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04*, CSE '09, 90–97. Washington, DC, USA: IEEE Computer Society.
28. F. Abel, I. Celik, G.-J. Houben, and P. Siehndel. Leveraging the semantics of tweets for adaptive faceted search on twitter. In *Proceedings of the 10th International Conference on The Semantic Web*, ISWC'11, pages 1–17. Springer-Verlag, 2011.
29. C. Lampe, and P. Resnick 2004. Slash(dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '04.