Identification of Useful User Comments in Social Media: A Case Study on Flickr Commons

Elaheh Momeni
University of Vienna
Währinger Strasse 29, A-1090
Vienna
momeni@cs.univie.ac.at

Ke Tao
Delft University of Technology
Mekelweg 4, 2628 CD, Delft
k.tao@tudelft.nl

Bernhard Haslhofer
Cornell University
Ithaca, NY 14850
bh392@cornell.edu

Geert-Jan Houben
Delft University of Technology
Mekelweg 4, 2628 CD, Delft
g.j.p.m.houben@tudelft.nl

ABSTRACT

Cultural institutions are increasingly opening up their repositories and contribute digital objects to social media platforms such as Flickr. In return they often receive user comments containing information that could be incorporated in their catalog records. Since judging the usefulness of a large number of user comments is a labor-intensive task, our aim is to provide automated support for filtering potentially useful social media comments on digital objects. In this paper, we discuss the notion of *usefulness* in the context of social media comments and compare it from end-users as well as expert-users perspectives. Then we present a machine-learning approach to automatically classify comments according to their usefulness. Our approach makes use of syntactic and semantic comment features and also considers user context. We present the results of an experiment we did on user comments received in six different Flickr Commons collections. They show that a few relatively straightforward features can be used to infer useful comments with up to 89% accuracy.

Categories and Subject Descriptors

H.3.m [**Information Storage and Retrieval**]: Miscellaneous

Keywords

User-generated Comment, Social Media, Usefulness Prediction, Flickr Commons

1. INTRODUCTION

Cultural institutions are increasingly contributing content to social media platforms to increase awareness and use of their collections. 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. Many of these platforms support annotation features, ranging from simple Like-button clicks to user-contributed full-text comments.

Comments can add supplemental information to existing digital resources, which might be interesting for other users. They may also contain factual information, such as names and places depicted on a photo, which is not available in existing metadata records. Such information can be gathered by institutions to improve descriptive metadata records and consequently to support efficient information retrieval and digital resource management [18, 10].

However, not all user-generated comments are useful for institutions and users. Users have different backgrounds, levels of expertise, and different intentions when contributing comments. As a result, the quality of user-generated comments varies from useful to useless and can even be abusive or off-topic. The two examples below, both taken from Flickr, illustrate the different nature of useful comments. The first one is useful and the second one is non-useful (labeled in our study [2]):

- (useful) "James Beauchamp Champ Clark (1850 - 1921) Speaker of the House (1911 - 1919), Missouri Senator. This picture may be from the Democratic Convention of 1912, where Clark was initially considered a frontrunner for the Presidential nomination. By the end, though, the nomination went to Woodrow Wilson."[3].
- (non-useful) "My great grandfather was an engineer at that time. I'd love to get a list of the names in that photo."[4]

Moderation by means of a dedicated human curator or forum administrator to identify useful comments is costly and time consuming and often not feasible given the potentially

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

[2]It is worth mentioning that comments which in our study are inferred as "non-useful" might be useful for other context and the term "useful" is a term that we use in our study.

[3]`http://www.flickr.com/photos/library_of_congress/2163461798/#comment72157603858091482`

[4]`http://www.flickr.com/photos/library_of_congress/2536790306/#comment72157629444651496`

JCDL'13, July 22–26, 2013, Indianapolis, Indiana, USA.

high number of comments and typically small number of staff members in cultural institutions. Therefore, we believe that automated approaches are needed that can segregate useful comments from non-useful ones.

Approaches to automatically estimate the usefulness of user-generated content (tags, tweets, and product reviews, etc) are gaining increasing attention [1, 17, 19]. However, each type of free-text user-generated content (e.g., product reviews, tweets, etc) has different characteristics compared to free-text comments and, furthermore, definitions and usage of useful user-generated content vary in different application scenarios. Some scenarios tend to prefer objective and informative content due to the promise of additional, potentially interesting information, while others see more value in the combination of subjective and objective content.

For the purposes of our current research, we are interested in identifying the possible characteristics of useful social media comments. The notion of usefulness is in itself very subjective and is largely determined by how different institutions apply it in specific scenarios. Therefore, in this paper, first we discuss the notion of *usefulness* with regard to comments that provide additional descriptive information of media objects from two perspectives – from users as well as experts perspectives. Then we show that comments which are useful for users are also useful for experts. To this end, the central contributions of this paper can be summarized as follows:

- *Identification of the characteristics of useful comments:* We gather a dataset of image comments from Flickr Commons[5] and collect users' and experts' usefulness judgments (by using a crowd-sourcing approach) to identify the usefulness of gathered comments. Then we 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.

Our findings reveal that a few relatively straightforward features can be used to infer the usefulness of comments. However, an analysis of the important features among different topic areas (place, person, and event) of media objects indicates that when inferring the usefulness of comments the influence of features varies slightly according to the topic areas of media objects. The major differences appear among the psychological content characteristics. Therefore, if prior to inferring usefulness we are able to determine the topic area of a media object, this helps to classify useful comments with higher accuracy. Thus, for a more accurate classification of useful comments, a model should be trained that takes into account the topic area of media objects.

[5]"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 you discover in The Commons on Flickr, either by adding tags or leaving comments." `www.flickr.com/commons`

We believe the comments which are useful for users and experts have potentially valuable information which may improve the retrieval and management applications. However, this is beyond the scope of this paper.

This paper 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 judgments and presents a series of usefulness classification experiments and evaluation of the derived features. Finally, we discuss and conclude our work in Section 5.

2. USEFUL SOCIAL MEDIA COMMENTS

The Oxford dictionary describes *usefulness* as "*a quality or fact of being able to be used for a particular or in several ways*". Thus, any characterization of usefulness depends on the institutional context and the application scenario. We found that the notion of usefulness has previously been discussed in two main research contexts: *assessing the quality of user-generated tags* and *assessing and ranking of user-generated free-text content.*

2.1 Assessing user-generated tags

Several works in the area of tagging and folksonomy research discuss the assessment of user-generated tags or the selection of tags that allow people to better describe their content. Sigurbjoernsson and van Zwol [19] propose approaches for the selection of useful tags by computing tags and URL co-occurrence patterns. They find that the tag frequency distribution follows a perfect power law, and they indicate that the mid section of this distribution contains the most interesting candidates for tag recommendation. Weinberger et al [22] define a measure of tag ambiguity, based on a weighted Kullback-Leibler (KL) divergence of tag distributions. Hall and Zarro [9] compare the abstracting and indexing practices of a semi-expert community (metadata creators for the digital library, ipl2) 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.

However, useful tags also provide descriptive information for objects despite the fact that user-generated comments have different characteristics from user-generated tags. User-generated comments are longer and have an informal structure and users can converse, express their subjective opinions and emotions, and describe informative useful information about a media resource.

2.2 Assessing user-generated free-text content

Assessing the quality of user-generated free-text information is critical in other domains such as question answering platforms, micro-blogging services (e.g.,Twitter), or product reviews.

Agichtein et al. [1] introduce a general graph-based classification framework for combining features from different sources of information in order to assess high-quality questions and answers in CQA (Community Question and Answer). Liu et al. [16] propose a method for predicting information seeker satisfaction in CQA and develop a variety of content, structure, and community-focused features

Features Groups	Ref	Short Description
Text statistics and syntactic features	[1, 4, 14, 6, 7, 17, 5]	Aggregate statistics extracted from the text such as length, readability, #token, etc
Semantic and topical features	[6, 7, 21, 22, 19, 13, 4, 15]	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	[17, 16, 21, 4, 6, 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

for this task. Harper et al. [11] propose an algorithm that reliably categorizes questions as informational or conversational. Castillo et al. [4] propose automatic methods for assessing 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. [6] develop methods for filtering and assessing the variety of sources found through social media by journalists. They take a human centered design approach for developing a system, which is informed by journalistic practices and knowledge of information production in events. Becker et al. [2] presented relevant Twitter content selection approaches and show that the centroid (as a centrality-based approach) emerged as the preferred way to select relevant tweets.

Furthermore, predicting the helpfulness of a product review (e.g., how many people have considered a particular product review helpful) is one of the related problems. Several approaches demonstrate 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 length of the review [14, 7], mixture of subjective and objective information [7], readability such as checking the number of spelling errors [7], and conformity (a review is evaluated as more helpful when its star rating is closer to the consensus star rating for the product) [14, 5]. Moreover, Lu et al [17] illustrate how the social features of reviewers can help the assessment process.

2.3 Taxonomy of Useful Social Media Comments

A preliminary exploration of these assessment and ranking techniques 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. In Table 1 we categorize these features and strategies into three feature groups.

In this paper, 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.

3. FEATURES ENGINEERING

With regard to available approaches and features for similar problems, explored in Section 2, we believe that there are some observable features in social media platforms themselves, that are useful to assess the usefulness of comments.

We provide an overview of the different features to characterize each comment in our feature collection. We select influential features from related work. These features are aligned with our assumptions of characteristics of useful comments. These include some features specific to the Flickr platform, but most of them are quite generic and can also be applied to other platforms. Our feature set is listed in Table 2. According to our study, in Section 2 we group potentially important features into three different groups:

Text Statistics and Syntactic Features (TS) This group captures surface-level identification of the usefulness and includes the following features:

- *Punctuation Mark*, counts the number of punctuation marks. We assume that comments, which contain a higher number of punctuation marks are less likely to be useful.

- *Text Statistics*, measures 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 features such as percentages of verbs, adverbs, etc. We use the LingPipe toolkit[6] to obtain the relative POS taggers. We assume that comments containing a higher number of words are likely to be useful [14, 7].

- *Linkage Variety*, counts the number of unique hyperlinks in a comment. We assume that the more links are contained in the comment, the more likely it is to be useful [4].

- *Informativeness*, measures the novelty of terms, t, of a comment, c, compared to other comments on the same object, calculated using:

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

  We assume that comments with higher informativeness score are more informative and, therefore, they are likely to be useful [21].

- *Readability*, measures how difficult the comment is to parse by using the Gunning fog index [8]. We assume that comments with a higher readability score are likely to be useful, because they are easier to parse for humans.

Semantic and Topical Features (ST) The semantics of a comment may increase or decrease the likelihood of a

[6] `http://alias-i.com/lingpipe/`

Feature	Short Description
Text Statistics and Syntactic Features (TS)	
Punctuation Mark	#punctuation marks
Text Statistics	#WC, #Verbs, #Adverb, average length of sentences (WPS)
Linkage Variety	#Hyperlinks in a comment
Informativeness	novelty of terms of a comment compared to other comments on the same objects
Readability	how difficult it is to parse the comment
Semantic and Topical Features (ST)	
Sentiment Polarity	positive and negative sentiment of content
Subjectivity Tone	the subjectivity or objectivity tone
Name Entities	#Name Entities
NE Types Variety	#distinct types of Name Entities that are mentioned in a comment
Psychological content characteristics	psychological dimensions of the content of a comment: Swear, Sadness, Anger, Family, Friends, Humans, Anxiety, Health, Sexuality, Home, Religion, Relativity, Leisure, Insight, Certainty, Tentativeness, Self-reference scores
Topical Conformity	the distance between the topics of a comment and the topics belonging to other comments on the same object
User Topic Entropy	the topical focus of the user via the entropy of topic distributions of the user
User and Social Features (US)	
User Linkage Behavior	#Hyperlinks posted by the user
User Conversational Behavior	#conversational comments posted by the user
User Activity	#Comment, #UploadedObject, #FavoriteObject scores
User Social Relation	#Contact, Prestige scores

Table 2: Overview of Features

comment being useful regardless of its text structure. Furthermore, this group includes standard topical model features, that measure the topical concentration of the author of a comment and topical distance of a comment compared to other comments on an object. This group includes the following features:

- *Named Entities*, counts number of named entities (NE) that are mentioned in a comment. We assume that the more named entities are mentioned in a comment, the more aspects of the object are covered. In this group, for NE related features we used GATE toolkit[7].
- *NE Types Variety*, counts distinct types of named entities (such as person, place, date, etc) that are mentioned in a comment. We assume that the more the types of entities are mentioned in a comment, the more aspects of the object are introduced. Therefore, the greater the diversity of concepts mentioned in a comment is, the more likely it is to be useful.
- *Subjectivity Tone*, the subjectivity or objectivity tone of a comment may impact the usefulness of the comment. We used Subjectivity Lexicon [23] to calculate subjectivity. We assume the subjectivity or objectivity tone may have influence on the usefulness of the comments [7].
- *Psychological content characteristics*, identifies psychological dimensions: leisure, anger, family, friends, humans, anxiety, sadness, sexuality, home, religion, relativity, affective process, and self-reference scores. We used LIWC [20] for analyzing psychological characteristics of the content of comments. LIWC identifies psychological dimensions in the text of comments such as self-reference scores (usage of "I"), anger score (e.g. hate, loathe, etc).
- *Topical Conformity*, measures the distance between the topics of a comment and the topics belonging to other comments on the same object. An LDA model (Latent Dirichlet Allocation [3]), was trained to handle features that depend on topic models. To train the LDA model we aggregated all the comments on photos in our database into an artificial photo document to infer topic distribution and chose the following hyper-parameters: $\alpha = 50/T$, $\beta = 0.01$ and T = 1,000. 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:

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

We assume that very high or very low topical conformity has an impact on predicting usefulness of the comment [22, 21].

- *User Topic Entropy*, measures the topical focus of an author via the entropy of topic distributions of a user, inferred via the comments she authored. To handle this feature we also trained an LDA model [3]. To train the LDA model we aggregated all the comments authored by one user in our database into an artificial user document to infer topic distribution by users and we chose the following hyper-parameters: $\alpha = 50/T$, $\beta = 0.01$ and T = 1,000. Then, we measured the distance topic distribution of each user. We define entropy of topic distribution of all comments authored by an author, a_i as:

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

Where 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*, measures the sentiment polarity of a comment as:

$$SenPolarity = \#(PosTerm) + \#(NegTerm)$$

Based on [7, 4], we assume the sentiment polarity may have influence in the usefulness of the comments.

User and Social Features (US) Different characteristics of users and their social context may increase or decrease the likelihood of their comments being useful. Due to limitations of access to this information, we apply a lightweight characterization of authors and their social contexts. This group includes the following features:

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

- *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. We assume that the comments by users that use other resources as references are more likely to be useful.

- *User Conversational Behavior*, counts comments which contain a @reply. We assume that users who write comments to converse with other users are less likely to write useful comments.

- *User Activity*, measures different activities completed by a user, such as: number of comments (counts the number of comments authored by the user), number of uploaded objects (counts the number of media objects uploaded by the user), and number of favorite objects (counts the number of media objects selected as favorite by the user). Based on [6, 4], we assume that the users' activities influence the usefulness of their comments.

- *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). We assume that users with a higher number of social interactions are more likely to write useful comments [17].

4. EXPERIMENTS

In this section we describe how we collect usefulness judgments for characterizing useful comments and then we report the results of different machine learning methods for training an inference method that attempts to automatically infer if a comment is useful or not. Furthermore, we evaluate the quality of the features for inferring the usefulness of a given topic.

4.1 Data Acquisition

First, we build a dataset from real world comments harvested from Flickr Commons. Second, we extract those comments that have attracted a response by experts of cultural institutes. Third, we use a crowd sourcing approach, set up a user study, and request people to state if they consider that a certain set of comments could be useful for them. In order to compare how users' perspectives of usefulness is similar to experts' perspectives we compare characteristics of useful user-judged and expert-judged comments.

Dataset: We compile our dataset from real world comments harvested from Flickr Commons. We crawl 32,132 comments written on 11,130 photos on Flickr Commons. The photos are selected from six different institutes. From each institute we select the photo-sets with the highest number of comments on their photos. Each of these photo-sets corresponds to what we call a topic. All topics - according to the titles of the selected photo sets - are selected with three broader topical focuses: events (e.g., Irish civil war, News in 1910, World War 1, etc), places (e.g., old New York, old Edinburgh, ruins in the Middle East, etc), and persons (e.g., Neil Armstrong,etc). Finally, in our dataset we have three sets of comments on three topics: place, person, and event. In one of the Library of congress photo sets (News in 1910), we recognize that it contains many photos about persons. Therefore, we separate photos which show only a photo of a person from other photos, which belonged to topics related to event, according to their titles. Furthermore, for training a classifier and analyzing users' features, we crawl all profile information of all users who wrote comments. Table 3 shows the summary statistics of the dataset.

Photoset	Topic	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
All		**32,132**	**11,130**	**6,934**

Table 3: Summary statistics of dataset

Collecting Expert Judgements for Defining Usefulness: With regard to comments written on photos of the Library of Congress (LOC), we recognize some of these comments are commented upon by the LOC experts[8]. In order to ensure that these comments are useful for LOC we asked LOC staff members what causes them to comment back. They confirmed that commenting back is one indicator of a useful comment: "all Flickr comments are being read by LOC staff. The vast majority of comments are 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 will know that their information is making a difference.". Based on these observations, first we crawled 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 used the crowd sourcing approach and we asked coders to assist us in defining relevant comments. We used 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 gathered amounted to 2,068 comments, which we presume to be considered as 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 context and the term "useful" is a term that we use in our study.

Collecting User Judgements for Defining Usefulness: In order to understand the characteristics of useful comments from user's perspective and to collect non-useful comments for training a classifier, we randomly selected 3,500 comments and crowd sourced to collect user usefulness judgments by using CrowdFlower.com. We asked coders to assist us in identifying useful comments. We showed each coder a comment and links to the related Flickr photo and asked them to answer four questions:

1. Text-box question, "How many Web links does the comment contain?"
2. Option-choice question, "Does the comment contain Web links?"
3. Text-box question, "Select 1 to 4 main keywords used in the comment."

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

4. Option-choice question, "Compared to the description provided by the uploader of the photo, is this comment useful for you to learn more about the content of the photo?"

In order to ensure the quality of the work by coders, for each comment we asked each coder to answer 3 objective questions, which can be computed automatically (questions 1 to 3) and one question (question 4) with regards to the usefulness of the comment. The first two questions are semantically the same but asked in two different ways. Inconsistency in answering the first two questions gave us the chance to exclude randomly selected answers. With regard to the third question we ask coders to enter text, which gives greater possibility to validate non-serious contributions. Therefore, inconsistency in selecting relevant keywords gave us the chance to exclude non-serious selected answers. Finally, the coders contributed to our task by answering the fourth question and the main question. For each comment we collected 3 independent judgments.

As result of this study, 1,345 of 3,500 comments (38.42%) received majority agreement on being useful. In order to examine the user agreement, we compute the level of the (inter-rater) agreement between coders based on Fleiss' Kappa for each comment. The mean Kappa score for all comments is above the score of 86%, indicating almost perfect agreement for the usefulness inference between coders.

Characteristics of Expert-judged vs User-judged Useful, and Non-Useful Comments In order to prepare a training-set for developing a usefulness classifier, first, we selected 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. Furthermore, in order to compare characteristics of user-judged with expert-judged useful comments we randomly selected 1000 expert-judged useful comments.

Second, we assess the mean values and standard deviations of each feature, which shown in Table 4. As expected, the average semantic and topical-based scores for comments which are judged as useful is different from those for 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. A comparison of the 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, 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, 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 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.

Third, we assess the mean values and standard deviations of each feature for expert-judged comments. Table 5 shows these values in comparison with user-judged useful comments in detail. This table shows the mean and standard deviations of almost all features from both datasets are in the same range. This result suggests that characteristics of user-judged comments are very similar to characteristics of expert-judged useful comments and therefore the non-useful comments (labeled in our study) can be assumed to be non-useful from both perspectives.

Features	Mean-U	STD-U	Mean-N	STD-N
Text Statistics and Syntactic Features (TL)				
Informativeness	14.50	21.91	05.02	06.37
Readibility	06.05	04.07	05.70	03.54
#Punctuation Marks	77.76	131.4	77.10	214.7
#WC	41.70	49.41	09.32	12.52
#WPS	15.63	10.99	06.36	06.50
#Verb	09.06	08.61	09.05	11.38
#Adverb	02.91	04.81	05.10	10.30
Linkage Variety	01.72	01.82	0.521	05.92
Semantic and Topical Features (ST)				
#Name Entities	03.62	05.33	0.466	0.956
NE Types Variety	01.39	01.07	00.36	00.58
Topical Conformity	01.34	01.67	01.07	01.10
Sentiment Polarity	01.62	03.75	29.26	32.77
Subjectivity Tone	0.151	0.160	0.910	0.750
Sadness	0.190	0.880	0.160	0.940
Insight	0.150	01.56	0.096	0.810
Anger	0.369	01.74	0.197	01.80
Family	0.460	01.63	0.126	01.40
Friends	0.060	0.950	0.130	02.98
Humans	0.590	01.93	0.840	03.64
Health & Body	0.790	02.41	01.93	07.02
Sexual	0.065	1.086	0.970	05.10
Religion	0.409	02.86	0.103	01.21
Leisure	01.30	02.99	0.460	02.51
Swear	0.058	0.087	0.198	0.682
Home	0.450	01.74	0.180	01.35
Relativity	12.86	09.18	06.14	09.87
Certainty	0.616	1.980	1.290	6.750
Tentative	01.79	03.65	01.21	03.98
Self-reference	01.02	2.587	02.27	05.42
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
#UploadedObject	20250	3869	1390	3134
#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 4: 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

4.2 Experimental Setup

First, we selected 1,000 useful comments with high agreements with regard to their usefulness and 1,000 comments with high agreements with regard to their non-usefulness from our labeled data. This amounted to a total of 2,000 comments for the training-set. Second, we evaluate the performance of four classifier algorithms for inferring the usefulness of comments: Logistic Regression, SVM, Naive Bayes and the decision-tree classifier J48. For analyzing the influence of the different sets of features on their performance, each classifier was set with all combinations of feature sets and they were evaluated against each other, using four measures: precision, recall, F1-measure (the harmonic mean between precision and recall) and area under the Receiver Operator Curve (ROC).

Results of Classification Evaluations Table 6 shows the performance of different sets of features and classification methods for predicting the usefulness of comments. It demonstrates the effectiveness of using semantic and user features for inferring useful comments. Training a classification model using semantic and user feature shows improved performance compared to the same models trained using

Features	Mean-U	STD-U	Mean-E	STD-E
Text Statistics and Syntactic Features (TL)				
Informativeness	14.50	21.91	15.36	25.47
Readibility	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 5: The comparison of the mean and standard deviation values of each feature between user-judged (U) and expert-judged (E) useful comments.

Feature-Sets	Classifier	Precision	Recall	F1	ROC
TL	LR	0.76	0.75	0.75	0.85
	SVM	0.75	0.74	0.74	074
	NB	0.74	0.71	0.71	0.77
	J48	0.79	0.79	0.79	0.84
ST	LR	0.83	0.82	0.82	0.91
	SVM	0.84	0.82	0.82	0.82
	NB	0.81	0.80	0.79	0.89
	J48	0.83	0.82	0.81	0.88
US	LR	0.71	0.69	0.69	0.80
	SVM	0.71	0.66	0.65	0.67
	NB	0.71	0.66	0.65	0.80
	J48	0.78	0.78	0.78	0.86
TL + ST	LR	0.85	0.85	0.85	0.89
	SVM	0.83	0.82	0.82	0.82
	NB	0.79	0.79	0.79	0.88
	J48	0.83	0.82	0.82	0.82
ST+ US	LR	0.85	0.85	0.85	0.93
	SVM	0.85	0.85	0.85	0.85
	NB	0.84	0.83	0.83	0.92
	J48	0.85	0.85	0.85	0.90
TL+ US	LR	0.83	0.83	0.83	0.90
	SVM	0.82	0.81	0.81	0.81
	NB	0.80	0.77	0.77	0.86
	J48	0.84	0.84	0.84	0.87
ALL	**LR**	0,88	0.90	**0.89**	**0.95**
	SVM	0.85	0.83	0.85	0.89
	NB	0.80	0.891	0.85	0.92
	J48	0.88	0.88	0.88	0.89

Table 6: Classification results of useful comments from non-useful comments using four classifier algorithms: Logistic Regression (LR), SVM, Naive Bayes (NB) and the decision-tree (J48) with all features-sets combinations

text-related features. By combining all the features we are able to achieve an F1 score of 0.89, coupled with high precision and recall when using the Logistic Regression classifier.

J48 classifier performs similarly, while SVM and Naive Bayes perform with lower accuracy. This demonstrates the poor precision and recall levels when using text features only. In such a case, each model fails to provide the same performance compared to when the semantic and user features are used.

Influence of Features on Usefulness Inference So far we have only analyzed the use of features grouped together. We now evaluate the quality of each individual feature for classifying useful comments.

In order to detect how the features were associated with the usefulness of comments, the coefficients of the Logistic Regression model were inspected from the selected best performing model (using all sets of features). A positive coefficient denotes a higher probability that the feature better correlates with usefulness. Table 7 shows detailed coefficient ranks. In addition to interpreting the statistically significant coefficients, we also ranked the features of the best performing feature group by using the Information Gain Ratio (IGR) as a ranking criterion. Table 7 shows the 20 top-ranked features, which are dominated by semantic features. Figure 1 shows the contributions by each of the top-10 features to classify usefulness in the training set, where the affective process (such as *Subjectivity Tone* and *Sentiment Polarity*) and Name Entities- related features of the comments seem to correlate strongly with inferring useful comments.

More precisely, coefficient ranks show that comments that expressing emotional and affective processes of the author (higher *Subjectivity Tone*, *Sentiment Polarity*, *Anger*, and *Sadness*) scores have a negative impact on the usefulness inference model) are more likely to be inferred as non-useful. Comments with offensive language (*Swear* score has a negative impact on the usefulness inference model) are more likely to be inferred as non-useful. Nevertheless, comments which have higher *Name Entities*, *NE Type Variety* and *Linkage* scores contain potentially interesting information and are likely to be inferred as useful. With regard to user features the *User Linkage Behavior* is a good indicator showing that users may diligently cite references for the information they provide. This increases reliability when inferring usefulness for such comments. Interestingly, a higher *User Topical Entropy* has a negative impact on the usefulness inference. Users with a higher entropy have a lower topical focus and therefore might write comments with less focus on the specific topic. Therefore, their comments are likely to be inferred as non-useful. A higher score of *Self-reference* and a higher *Conversational Behavior* score also have a negative impact. This suggests that users, who mostly use systems to converse and describe their personal experiences do not write useful comments. 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.

A high readability score does not have an important impact on usefulness inference. This is the case for many useful comments and is due to the fact that comments which are longer and contain more complex words are less readable based on the Gunning fog score [8]. The usage of insight terms (such as "think", "know", "consider") shows a positive correlation with usefulness. Furthermore, the usage of certainty terms (such as "always", "never" etc) has a negative impact on the model. This might be due to the fact that users who are assertive and express certainty tend to be seen as more subjective and less analytical. In contrast, the usage of tentative terms (such as "maybe", "perhaps", "guess", etc) shows that authors do not make any claims as to the cor-

Rank	Features	Coefficient
1	ST-Subjectivity Tone	-3.828
2	ST-Sentiment Polarity	-1.157
3	ST-NE Types Variety	0.550
4	US-User Linkage Behavior	0.025
5	ST-#Name Entities	0.211
6	ST-Self-reference	-0.148
7	TS-WPS	0.031
8	ST-User Topic Entropy	-0.049
9	ST-Insight	0.049
10	ST-Swear	-0.045
11	TS-Linkage	0.173
12	US-User Conversational	-0.023
13	ST-Certainty	-0.032
14	TS-Future Verb	-0.043
15	TS-Impersonal-pronoun	0.025
16	US-Prestige score	0.060
17	ST-Religion	0.089
18	ST-Sadness	-0.075
19	ST-Family	0.016
20	ST-Relativity	-0.006

Table 7: Top-20 features and their coefficient, derived from Logistic Regression model. Ranks are based on IGR ranking

rectness or certainty of their comments and such comments are likely to be inferred as useful.

In order to observe the effects of iteratively increasing features, and the impact on classification performance, our next experiment explored the effects of training a classification model using only the top-ranked features. We selected the Logistic Regression classifier for training - based on its optimum performance during the model selection phase - and trained the classifier using the training split from each dataset. Figure 2 shows the results from our experiments, where at lower levels of ranked features we observe similar levels of performance. As we included more features within our classification model, we observe improvements in F1 scores. As we include the lower-ranked features, our plots show only a minor increase in 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. However, Semantic and Topical features play important roles.

Influence of Topics on Usefulness Inference In all reported results so far, we have considered the entire set of topics. Therefore, we investigate to what extent certain topic characteristics play a role with regard to usefulness inference and to what extent those differences lead to a change in the inference models.

From our training data we creat three different test-sets of comments related to each area of topics (person, place, and event) and then, for each set we predicted the usefulness of comments using two models: first, the usefulness classification model, which was trained with regard to the area of topic, second, a general usefulness classification model for all topics (the model, that does not take into consideration the topic of media objects). Furthermore, we perform three Pearson's Chi-squared tests between the two predicted results for each topic from different models. Our findings from Table 8 show that, despite the F1 and ROC measures for both prediction results are slightly different for all topics, the p-value means are evidence ($p < 0.01$) for topics related to person and place, that the two predicted samples come from different distributions. These results suggest, if prior to inferring usefulness we are able to determine the topic area

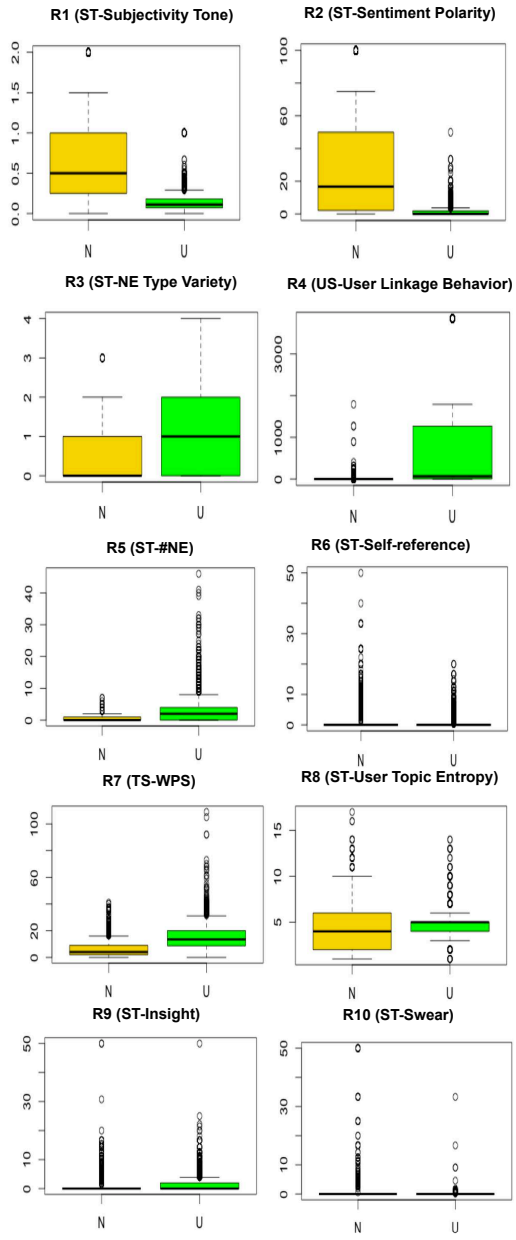

Figure 1: Box plots for the top 10 features according to IGR and Coefficient ranking. Yellow boxes (class N, left) represent non-useful comments, green boxes (class U, right) represent useful comments

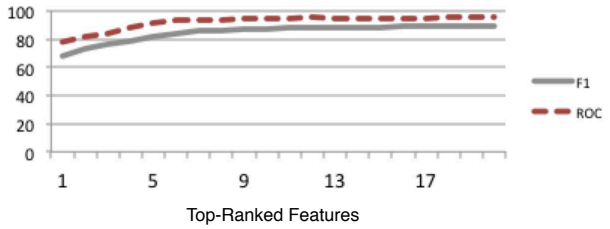

Figure 2: Performance results of classification using top-20 features

Model	Person		Place		Event	
	All	Person	All	Place	All	Event
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

Table 8: Evaluation results of classification models for different topics. All is the model, which does not take into consideration the topic of media objects. The star next to the F1 means that there is evidence ($p < 0.01$) that the two predicted samples come from different distributions.

of a media object, this helps to classify useful comments with higher accuracy.

We investigate the difference between the models derived for each of the topics. Table 9, shows detailed coefficient ranks for different models. An analysis of the most important features among different areas of topics (place, person, and event) shows some differences. More precisely, coefficient ranks show that comments related to the topics person and event express the users' emotional and affective processes, which are more likely to be inferred as non-useful. An analysis of the *Subjectivity Tone* in different topics shows that the *Subjectivity Tone* for person-related topics is also higher than for other topics. An analysis of the *Swear* score among different topics shows that the *Swear* score for comments related to person is the most negative one. For comments on event, the *Swear* score is more negative than for topics related to place. This concurs with our observation that users express more emotion and may use more offensive language when writing comments about topics related to persons and events. Such comments are more likely to be inferred as non-useful.

A topic related to an event is often also related to a person, a place, or both. Therefore, the coefficient ranks are influenced by the two other topics. For example, the *Relativity* score which includes physical place and motion has a positive impact on places and events, while it has a negative impact on topics related to persons. It is interesting to note that for topics related to place, *Relativity* scores have a positive impact on the model. *Friend* and *Family* scores have a negative impact on the model. This might be because that people describe different physical phenomena and motion processes on this topic, which may be seen as useful information by other users. Instead, giving information about friends and family is non-useful for other users. With regard to topics related to person, *Family, Health & Body* scores have a positive impact on the model. This might be because that people describe more about various health and body aspects of a person on these topics. Furthermore, they describe the background of family members of the target person. This information may be useful information for other people.

Our results indicate that, for a more accurate classification of useful comments, a classification model should take into account the topic of media objects.

5. DISCUSSION

We conducted an analysis of user-generated comments on media objects of different museums and libraries to shed some light on the characteristics of useful comments and to identify the important key features of comments for inferring usefulness. We analyzed three different sets of features: text

Rank	Features	Place	Person	Event
1	**ST-Subjectivity Tone**	-4.271	-6.228	-3.406
2	**ST-Sentiment Polarity**	-0.157	-0.223	-0.647
3	**ST-NE Types Variety**	-0.138	0.113	0.776
4	**US-User Linkage Behavior**	0.046	0.003	0.002
5	**ST-#Name Entities**	0.203	0.109	0.201
6	**ST-Self-refrence**	-0.161	-0.136	-0.177
7	**TS-WPS**	0.055	0.029	0.016
8	**ST-User Topic Entropy**	-0.112	-0.302	-0.059
9	**ST-Insight**	-0.124	0.081	0.064
10	**ST-Swear**	-0.005	-90.42	-3.363
11	**TS-Linkage**	0.084	3.028	0.610
12	**US-User Conversational**	-0.086	-0.086	-0.066
13	**ST-Certainty**	0.110	0.042	-0.054
14	**TS-Future Verb**	-0.071	-0.027	-0.027
15	**TS-Impersonal-pronoun**	-0.052	-0.040	-0.042
16	**US-Prestige score**	0.162	0.005	0.070
17	**ST-Religion**	-0.361	0.322	0.089
18	**ST-Sadness**	-0.110	-0.403	-0.038
19	**ST-Family**	-0.196	1.111	-0.004
20	**ST-Relativity**	0.163	-0.160	0.029

Table 9: Coefficient for features of models derived for different topics with regard to usefulness inference

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. This suggests that comments which 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) are more likely to be inferred as useful. Therefore, we suggest that a commenting system should give and motivate users to define references [12]. This adds unambiguous users-verified concept references to social media comments, which in turn has a positive impact on the usefulness of comments.

An analysis of features related to users suggests that by leveraging users' previous activities we may be able to increase the probability for inferring the usefulness of a comment. Therefore, we suggest that by designing a commenting service, designers should take this into account when designing users' profile pages.

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. Users express more emotion and may use more offensive language when writing comments about topics related to persons and events. Such comments are more likely to be inferred as non-useful. For topics related to place, people describe more physical phenomena and motion processes on this topic, which may be seen as useful information by other users. On the other hand, giving information about family is non-useful for other users. In contrast, for 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. Therefore, if prior to inferring usefulness we are able to determine the topic area of a media object, this helps to classify useful comments with higher accuracy.

We believe that our results may also apply to other social media platforms. However, we believe the influence of features may vary according to the commenting cultures of platforms. Therefore, in future work, we will further explore the impacts of features on other platforms. We will also explore, different aspects of topics of media objects (such as temporal, polarization, etc), which might have influence on building better classification models.

6. 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.

7. 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] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *the Journal of machine Learning res*, 2003.

[4] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *the 20th international conference*, WWW, 2011.

[5] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, and L. Lee. How opinions are received by online communities: a case study on amazon.com helpfulness votes. In *Proceedings of the 18th international conference on World wide web*, WWW '09, 2009.

[6] 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.

[7] 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.

[8] R. Gunning. *The Technique of Clear Writing*. McGraw-Hill, New York, 1952.

[9] 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.

[10] 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.

[11] F. M. Harper, D. Moy, and J. A. Konstan. Facts or friends?: distinguishing informational and conversational questions in social q&a sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2009.

[12] B. Haslhofer, W. Robitza, C. Lagoze, and F. Guimbretiere. Semantic tagging on historical maps. In *ACM Web Science 2013*, Paris, France, May 2013. ACM.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. 2010.

[21] 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.

[22] 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.

[23] 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.