Leveraging Semantic Facets for Adaptive Ranking of Social Comments

Elaheh Momeni
University of Vienna
Vienna, Austria
elaheh.momeni-ortner@univie.ac.at

Reza Rawassizadeh
Dartmouth College
Hanover, NH
rrawassizadeh@acm.org

Eytan Adar
University of Michigan
Ann Arbor, MI
eadar@umich.edu

ABSTRACT

An essential part of the social media ecosystem is user-generated comments. However, not all comments are useful to all people as both authors of comments and readers have different intentions and perspectives. Consequently, the development of automated approaches for the ranking of comments and the optimization of viewers' interaction experiences are becoming increasingly important. This work proposes an adaptive faceted ranking framework which enriches comments along multiple semantic facets (e.g., subjectivity, informativeness, and topics), thus enabling users to explore different facets and select combinations of facets in order to extract and rank comments that match their interests. A prototype implementation of the framework has been developed which allows us to evaluate different ranking strategies of the proposed framework. We find that adaptive faceted ranking shows significant improvements over prevalent ranking methods which are utilized by many platforms such as YouTube or The Economist. We observe substantial improvements in user experience when enriching each element of a comment along multiple explicit semantic facets rather than in a single topic or subjective facets.

CCS CONCEPTS

•Information systems →Information retrieval;

KEYWORDS

Adaptive Ranking; Social Comment; Semantic Facets

ACM Reference format:
Elaheh Momeni, Reza Rawassizadeh, and Eytan Adar. 2017. Leveraging Semantic Facets for Adaptive Ranking of Social Comments. In *Proceedings of ICMR '17, Bucharest, Romania, June 06-09, 2017,* 9 pages.
DOI: http://dx.doi.org/10.1145/3078971.3079004

1 INTRODUCTION

Online social media systems provide commenting features to support augmentation of social objects — ranging from YouTube videos, Flickr images, SoundCloud audio fragments to more classic news articles. Many platforms also (such as Facebook) allow commenting on any type of media object by embedding some javascript. Good comments contribute multiple perspectives and observations, answer questions, form hypotheses, and generally contribute to the development of the "social media object". Unfortunately, most comment presentation systems only provide simple temporal streams. These contain a diversity of focus, usefulness, and quality (with many comments being abusive or off-topic). Due to the substantial number of comments, without a mechanism for end-users to unravel comment streams and identify those likely to be of interest, most end-users tend to become inundated by the number of comments and disillusioned by their experience. End-users may even stop their participation and contributions.

The usual method to provide comment recommendation support is simply to ask end-users to evaluate contributions by others. This wisdom-of-the-crowd approach (crowd-based ranking) allows all users to vote on or rate comments. However, Liu et al. [10] show that voting is influenced by a number of factors (including the "rich get richer" phenomenon) that may distort accuracy. An alternative method has relied on classification to determine the usefulness of comments, with the training data being a set of majority-agreed labeled comments [15]. This approach avoids some of the biases that arise from crowd voting, but removes control from the end-user and thus does not permit individual users to adapt the ranking according to their preferences.

In this work, we propose and evaluate an adaptive faceted ranking framework for social media comments that addresses this problem and allows end users to identify comments of interest to them. We investigate how to design appropriate faceted ranking strategies for commenting systems and analyse the impact of different types of facets on the ranking performance.

This enables comments to be accessed and ordered in multiple ways rather than in a single topic order [1, 2], which was demonstrated in the past to be more in line with users' desires and in addition it overcomes the cold-start problem. This framework enables users to explore different semantic facets and select combinations of facets in order to rank and extract comments that match their interests or are relevant to selected facets.

For example, a YouTube video, *"Steve Jobs' 2005 Stanford Commencement Address"*[1] contains more than 20,000 comments. Authors of the comments have discussed different aspects of Steve Jobs's life subjectively and objectively. The commenting system of YouTube currently provides 20 comments per page using reverse-chronological ranking or top recommended comments using votes of the other users (crowd-based ranking). So, if a user is interested to read informative comments, available ranking features do not satisfy the information need of the user as many subjective and personal opinions comments are positioned at the top of these ranking lists.

[1] https://www.youtube.com/watch?v=UF8uR6Z6KLc

ICMR '17, Bucharest, Romania
© 2017 ACM. 978-1-4503-4701-3/17/06…$15.00
DOI: http://dx.doi.org/10.1145/3078971.3079004

Instead, our ranking framework proposes a set of semantic facets such as topics, informativeness, subjective opinions, enabling users to select combinations of facets to suit users' needs and rank comments with regard to their interests or relevance to selected facets. Furthermore, users by exploring various facets can have a better overview of the content discussed among authors and discover interesting discussions. For example, our framework also recommends as a type of facet a list of topics discussed among comments such as "Japan" or "Bill Gates" for the example video. So, users discover that there are some discussions about Steve Job's life in Japan and they can select two facets, "Japan" and "Informative" to read informative comments related to this topic.

To investigate the utility of our framework, we implemented a Web-based prototype [11], which enables end-users to interactively rank comment feeds using adaptive faceted ranking, thus allowing us to evaluate their experiences. The framework also enables users to vote whether comments match their interests or are relevant to their selections of facets to assess the ranking method's effectiveness. To investigate our framework, we experimented with two different classes of media objects – videos and news articles – using comments harvested from the popular online media platforms, YouTube (youtube.com) and The Economist (economist.com). We investigated two platforms, which are dynamic, deal with large content, and handle millions of comments, to demonstrate the scalability of our approach. Then, we present two studies: (1) a study on the effectiveness of adaptive faceted ranking and facet selection strategies and (2) a study on semantic enrichment and facet extraction from comments.

Accordingly, this work makes the following contributions:

- It demonstrates operationalisation of usefulness through strategies based on semantic enrichment and novel facet extraction and selection from comments.
- It reveals a framework and its building blocks that allow various faceted ranking strategies for content exploration of commenting systems, thus leveraging extracted facets to directly enable end-users to rank social media comments based on their preferences and interests.
- It presents an evaluation environment to compare the proposed framework with prevalent default methods utilized by many platforms. It also investigates the effectiveness of strategies for the selection of different types of facets.

In this work, we present a number of promising findings. We find that our proposed system shows significant improvements over prevalent default methods utilized by many platforms, such as reverse-chronological ranking or crowd-based ranking for enabling end-users to identify comments of interest to them. More precisely, for Economist comments, considering the interestingness dimension, adaptive faceted ranking reaches an effectiveness level of 0.85 in MAP (Mean Average Precision), while the standard reverse-chronological ranking and crowd-based ranking only reach 0.40 and 0.50 MAP respectively. For YouTube comments, adaptive faceted ranking reaches an effectiveness level of 0.71 in MAP, while the standard reverse-chronological ranking and crowd-based ranking only reach 0.46 and 0.26 MAP respectively. Next, we find that not all types of facets are equally useful, but instead that effectiveness varies according to three types of facets selected for this study: Topic-related facets, Subjective facets, and Objective facets.

We believe that this is the first study that investigates the effectiveness of different ranking methods on social media comments, explores appropriate faceted ranking strategies on commenting systems, and proposes different methods to extract various semantic facets.

2 RELATED WORK

Existing approaches for ranking of user-generated content as typically found in social media systems generally adopt one of three approaches:

Machine-based ranking approaches Some available approaches employ machine-learning methods to assess and rank content with regard to the defined "value". Siersdorfer et al. [15] and Momeni et al. [13] propose a classifier for the curation of useful comments on social media objects such as YouTube videos. A method for assessing the quality and credibility of a given set of tweets was proposed by Castillo et al. [4]. However, machine-based approaches which are trained as classifiers to rank comments are based on a set of majority-agreed labeled comments. This avoids some of the biases that emerge due to crowd-based voting, but removes control from end-users and thus does not permit individual requester to adapt ranking of comments based on their preferences [12]. This work proposes an adaptive faceted ranking framework that tackles this problem and allows end-users to identify comments of interest to them.

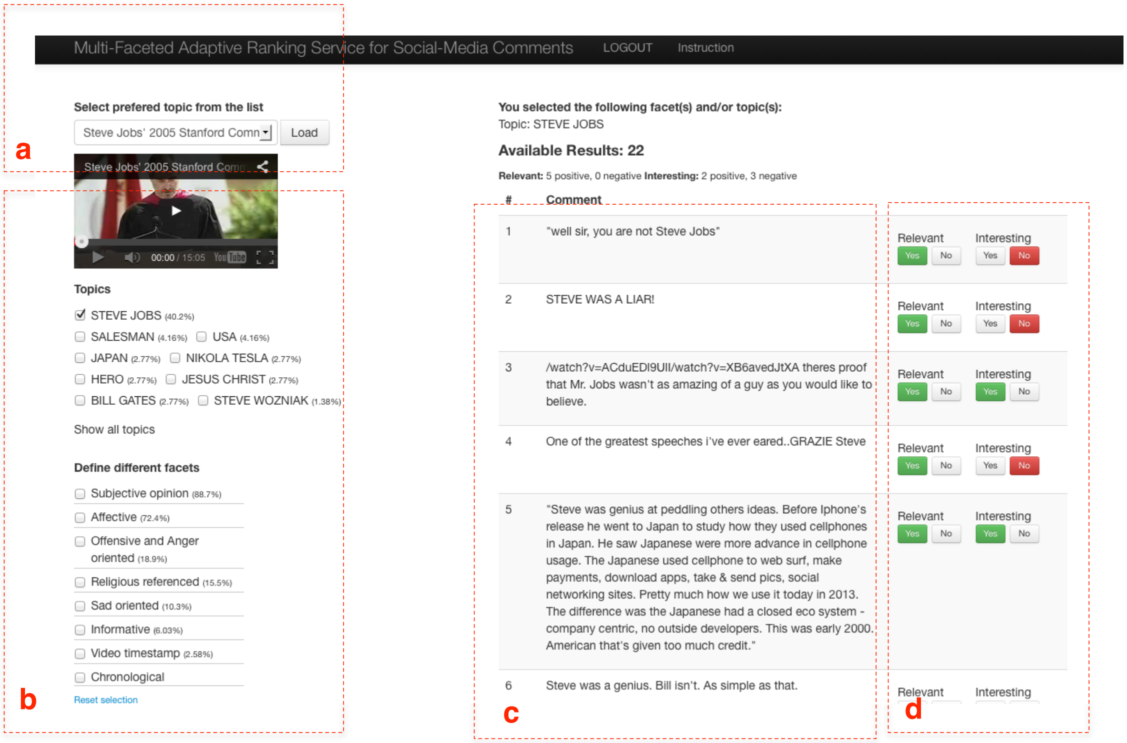
Personalized approaches For postings in micro-blogging platforms (such as Twitter), ranking and filtering methods based on collaborative filtering techniques are proposed by Hong et al., Das Sarma et al, and Paek et al. [5, 6, 14]. Furthermore, for posting in online forums, Lampe et al. [9] recommended that for ranking comments, patterns recognized by setting filters of users can be used to minimize the cost of settings for other users. These approaches provide personalized ranking of content, however they do not enable end-users to interact with the system and adapt the ranking with regard to their preferences.

Topic-based browsing approaches Other adaptive ranking solutions have focused on topic-based browsing which groups the comments into coherent topics and creates interfaces that enable users to browse their feeds more efficiently. Abel et al. [1] propose strategies for inferring facets and facet values on Twitter by enriching the semantics of individual Twitter messages. Topic modeling based methods (both on users and content) feature prominently in this space. Bernstein et al. [2] propose a browsable tag cloud of all the topics in a user's feed, allowing users to more easily find tweets related 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. We propose instead that a system that combines higher level features alongside topic classification is desirable.

3 ADAPTIVE FACETED RANKING

This work proposes a framework for enabling adaptive faceted ranking on comments attached to social media objects (such as comments


Multi-Faceted Adaptive Ranking Service for Social-Media Comments
LOGOUT
Instruction
Load
Show all topics
Reset selection
a
b
c
d

Select prefered topic from the list

Steve Jobs' 2005 Stanford Comn

Steve Jobs' 2005 Stanford Comm 00:00 / 15:05

Topics

☑ STEVE JOBS (40.2%)
□ SALESMAN (4.16%) □ USA (4.16%)
□ JAPAN (2.77%) □ NIKOLA TESLA (2.77%)
□ HERO (2.77%) □ JESUS CHRIST (2.77%)
□ BILL GATES (2.77%) □ STEVE WOZNIAK (1.38%)

Define different facets

□ Subjective opinion (88.7%)
□ Affective (72.4%)
□ Offensive and Anger oriented (18.9%)
□ Religious referenced (15.5%)
□ Sad oriented (10.3%)
□ Informative (6.03%)
□ Video timestamp (2.58%)
□ Chronological

You selected the following facet(s) and/or topic(s):
Topic: STEVE JOBS

Available Results: 22

Relevant: 5 positive, 0 negative **Interesting:** 2 positive, 3 negative

#	Comment	Relevant	Interesting
1	"well sir, you are not Steve Jobs"	[✓ Yes] No	Yes [✓ No]
2	STEVE WAS A LIAR!	[✓ Yes] No	Yes [✓ No]
3	/watch?v=ACduEDI9UlI/watch?v=XB6avedJtXA theres proof that Mr. Jobs wasn't as amazing of a guy as you would like to believe.	[✓ Yes] No	[✓ Yes] No
4	One of the greatest speeches i've ever eared..GRAZIE Steve	[✓ Yes] No	Yes [✓ No]
5	"Steve was genius at peddling others ideas. Before Iphone's release he went to Japan to study how they used cellphones in Japan. He saw Japanese were more advance in cellphone usage. The Japanese used cellphone to web surf, make payments, download apps, take & send pics, social networking sites. Pretty much how we use it today in 2013. The difference was the Japanese had a closed eco system - company centric, no outside developers. This was early 2000. American that's given too much credit."	[✓ Yes] No	[✓ Yes] No
6	Steve was a genius. Bill isn't. As simple as that.		

Figure 1: User interface of our developed prototype

on an online video or a news article). Our goal is to help users explore the comment stream by offering facilities to adapt the ranking of comments on the fly according to users' preferences.

User-Experience of Adaptive Faceted Ranking: To investigate the utility of our framework, we built a prototype Web-based interface [11][2] for our proposed ranking framework (Figure 1). With this prototype, end-users can experience the faceted adaptive ranking of social media comments and we can evaluate and track the end-users' interactions. For example, Alice as a user of the system wants to explore *informative* comments about Steve Jobs. She pulls up the interface, sees at the top of the interface (Figure 1-a) a drill-down list and selects a media object (an online video or news article). This triggers the system to crawl all comments related to the media object, enrich each comment along different semantic facets (such as *Subjective Opinion*, *Informative*, and *Topics*), dynamically extracts a set of facets for a set of comments, and recommends a list of facets on the left side of the interface (Figure 1-b). Additionally, in order to enable users to quickly acquire information about recommended facets, the system also shows percentages of related comments to the facet on the media object beside each facet. Alice now sees a list of facets (topic-based facets at the top and feature-based facets in the lower part). She explores facets and in addition to *"Informative"* facet, she can also see *"Nikola Tesla"* and other topics as facets. So, she selects a combination of facets based on her preferences. This triggers the system to present a ranked list of comments with high informative scores related to Steve Jobs on the right side of the interface (Figure 1 part-c). Also, the framework provides the conversation thread of each comment to Alice, helping her to understand the context of a comment. Finally, Alice can browse ranked comments and vote whether the comments match her interests and are relevant to her selection of facets (Figure 1 part-d).

For this prototype, three types of facets were implemented: **Topic-related facets**, topics discussed within comments on a media object. **Subjective facets**, such as comments with subjective tone, highly affective language, offensive and anger oriented, and sad oriented. **Objective facets**, such as informative, video timestamp, and religion referenced. Table 2 shows an overview of a set of semantic facets developed for this prototype and Table 4 shows some examples of comments for different facets. Related works on available approaches for analyzing free-text user-generated content in social media encourage the usage of these facets. Table 2 shows related work for each facet.

It should be noted that the set of facets explored in this work comprises a minimum set which demonstrates the effectiveness of using adaptive faceted ranking. However, it is not necessarily a complete set of useful facets. Some facets we developed include emotional facets such as "Subjective opinion" and "Affective" or ideological beliefs such as "Religious referenced", but other emotional facets such as anxiety or other ideological beliefs such as political orientation would be worth considering. Furthermore, we designed the presentation of facets as simple check boxes, our goal being to analyse the effectiveness of faceted ranking rather than design of the

[2] The development of the backend of the interface uses the REST style, permitting the interface to be easily integrated in any social media platform.

Feature	Description
Text Statistics	#Words, #Verbs, #Adjective, #Adverb, average length of sentences
Punctuation Marks	#Punctuation marks
#Named Entities	#Named Entities
NE Types Variety	#Distinct types of named entities
Readability	How difficult it is to parse the comment using Gunning's Fog Index
Psychological Characteristics	psychological dimensions of content of comment relies on the LIWC Scores[16]: Swear, Sadness, Anger, Anxiety, Religion, Positive Sentiment, and Negative Sentiment scores
Term Conformity	Distance between terms, *t*, of a comment, *c*, compared to other comments on the same media object, calculated using:$\Sigma_{t \in c} tfidf(t, c)$
Linkage Variety	#Hyperlinks in a comment
Subjectivity Tone	Measures the subjectivity degree of a comment. The Subjectivity Lexicon [18] is used to calculate subjectivity.

Table 1: Overview of basic features

facets. Nevertheless, we believe that different designs and orderings of facets may have a significant impact on the user interaction and experience, although this investigation is beyond the scope of this work.

System Architecture:

The proposed framework consists of four main components:

1. Semantic Enrichment Component: As comments do not explicitly feature facets, this component enriches each comment along a set of semantic facets. It first extracts a set of basic features considering the text of the comment (such as text statistics, subjectivity tone, etc) and using some lexicons. Table 1 shows an overview of a set of basic features. Second, it enriches each comment along various semantic facets. As shown in Table 2, proposed facets by framework are categorized into three broad types of facets. We developed facets using three methods: (1) Representing direct value of basic features, extracted from texts of comments (such as "Subjective opinion" or "Religious referenced"). (2) Calculating a synthesis value of some basic features' values (such as "Offensive and Anger oriented", which is calculated using average values of "Anger" and "Swear" features) (3) Assessing a value, exploiting trained classification models and basic features based on standard classification approaches (for example, "Informative" facet uses the logistic regression classifier and a subset of basic features). As the classification model for the trained facet is developed using binary classes (whether a comment is informative or not), we used prediction confidence scores for ranking comments. Considering topic-related facet - by experimenting with a number of topic labelling approaches (see Study 2), we found that named entities are useful proxies for the topics of the comments.

2. Facet Extraction and Selection Component: It is important to recommend the user most important and relevant set of facets, therefore this component operates on semantically enriched comments, extracts a set of facets, selects a list of proposed facets dynamically, and recommends the list of selected facets to the user. For example if a set of comments contains 100 topics as facets and some topics are relevant to only a very small set of comments (only relevant to 2 comments), then this component does not recommend these topics as facets. For a dynamic selection of facets, we chose Greedy Count a core algorithm. The algorithm ranks the facets for selection according to the number of top-n comments in the ranked result list and is similar to the Most Frequent heuristic selection used by [8].

3. Ranking Component: This component enables an individual user to explore proposed facets and select a combination of facets and then it ranks comments accordingly. For every ranking, the system returns a count of how many matching comments were tagged with each value within each combination of facet. Multiple selections of facets are treated as an 'and' rather than an 'or'. This means that the values of facets are combined conjunctively for the ranking of comments.

Facet	Description
Subjective Facets (SF)	
Subjective opinion	Represents direct score of the "Subjectivity Tone" feature (Encouraged Wagner et al. [17]).
Offensive and angry	Calculated using average values of "Anger" and "Swear" features (Encouraged by Momeni et al. [13]).
Sad	Calculated using average values of "Sadness" and "Negative emotion" features (Encouraged by Wagner et al.et al. [17]).
Affective	Calculated using $\frac{\#Positive\ emotion+\#Negative\ emotion}{\#Words}$ (Encouraged by Castillo et al. [4]
Objective Facets (OF)	
Informativeness	The extent to which an author provides useful information. Exploits a trained classification model using "Readability", "#Named Entities", "Linkage Variety", "Subjectivity Tone", and "'Term Conformity" features (Encouraged dby Momeni et al. [13])).
Religious referenced	The extent to which an author employs religion-oriented words such as "mosque" or "church". Represents direct score of the "Religion" feature (Encouraged by Momeni et al. [13]).
Video timestamp	Particular facet for video. The extent to which a comment points to a part of a video. Extracting entities related to time using extracted named entities and NE-related features (Encouraged by Momeni et al. [13]).
Topic Facets (TF)	
Topic	A multi facet which shows various topics discussed among comments using extracted named entities and NE-related features (Encouraged by Bernstein et al. [2])

Table 2: Overview of our samples proposed facets.

4. Feedback Collector: The goal of this component is to enable end-users to provide implicit and explicit feedback. This feedback facilitates the evaluation of different strategies related to various facet types. Implicit activities of users in the system such as exploration and selection of facets can be used as implicit feedback, and, furthermore, the system provides users with the chance to vote (explicit feedback) if a comment is "relevant" and/or "interesting". We have used these two scores to capture both the specific relevance of the comments to the facets and users' interests. This framing, we believe, is more interpretable from end-users' perspectives (as compared to "usefulness") and is also more nuanced than an up- or down- vote or simple score. Notably, a relevant comment is not necessarily interesting (and vice versa).

4 EXPERIMENTS

We now turn to evaluation of the framework. Specifically, our first study investigates answers to the following two questions: How well does adaptive faceted ranking compare to the prevalent ranking methods, such as reverse-chronological or crowd-based ranking? Which type of facets perform best in the ranking process and what are effective strategies for selecting and building facets? The second study evaluates particular aspects of the proposed framework such as: How accurate is semantic enrichment and facet extraction? Considering the topics discussed within comments as one type of facet, which topic-identification algorithm is most appropriate for comments?

Datasets: We investigate our experiments on two datasets from real-world comments harvested from the popular online media platforms, YouTube and The Economist. These provide free-text comments on different types of media objects (videos and news articles) from a variety of people with different backgrounds and intentions.

Dataset 1-YouTube Dataset: The YouTube dataset was provided by Momeni et al. [13] and extended by this work for the evaluation of the proposed ranking framework with two prevalent ranking methods: reverse-chronological and crowd-based ranking. The history timeline provided by About.com was used to identify popular topics of different time periods. Examples of topics are the "Irish civil war" and "1936 Olympics" as events,"old New York" and "old Edinburgh" as places, and "Neil Armstrong" and "Princess Diana" as people. Next, via the YouTube API a search was conducted with each of these topics. Those videos with the highest number of comments or a high number of views (and at least 100 comments) were selected. In total, 308 videos were included in this data set. From those, $91,778$ comments were crawled. For each video, the latest $1,000$ comments were crawled by using the reverse-chronological option offered by the YouTube API, permitting an evaluation of the effectiveness of reverse-chronological ranking. Furthermore, in this work, the top 50 crowd-based comments were crawled for each video by using the "Top Comments" option offered by the YouTube user-interface.

Dataset 2-Economist Dataset: For the second dataset, a set of comments occurring on the GD ("Daily Chart") of the The Economist platform between November 2011 to October 2013 were collected [7], resulting in a dataset of 15,870 comments across 546 posts containing one or more comments. GD is a publicly accessible blog that is part of The Economist online. Each workday the blog publishes "charts, maps and infographics", which are mostly static but sometimes include interactive visualisations. Posts also include a one or two paragraph text article contextualising the graphic. Collected comments are pseudoanonymous and appear below the visualisation in paged blocks of twenty. Comments are sorted in reverse chronological order by default, but we also collected the order of comments using recommendation scores, provided by the platform itself.

4.1 Study 1: Effectiveness of Adaptive Faceted Ranking Strategies

In this study, we conducted a laboratory study in order to, firstly, evaluate the effectiveness of the faceted framework and, secondly, investigate the effectiveness of strategies for the selection of different types of facets. We gave users a short period of time to explore a comment feed using three different ranking interfaces, and then asked them to provide us feedback. In this study, beside our approach (**Adaptive Faceted, AF**) we consider two popular default ranking methods: (1) **Reverse-Chronological (RC):** This method orders comments according to the time of posting (e.g., called *"Newest first"* by YouTube), and (2) **Crowd-Based (CB):** This method ranks comments by enabling each user to vote and judge the contributions of others, such as thumbs-up/thumbs-down or voting style. Accordingly, the platforms rank comments using these votes (e.g., called *"Top Comments"* by YouTube).

Participants: We recruited study participants from two large universities (names omitted for anonymity) by distributing calls for participation through internal science mailing lists. From the respondents, we randomly selected 45 subjects. Participant ages ranged from 20 to 57 with a median of 29. 26 subjects are students and 19 are other professionals such as administrative staff, lecturers, etc. Participants received a gift voucher as renumeration for evaluating the system.

Method: The study utilized a within-subjects design to directly compare adaptive faceted ranking to the prevalent ranking methods. To ensure that any observed differences were due to the propose interface and not to nuances of layout or colouring, users in same interface saw ranked list of comment using reverse-chronological and crowd-based methods. Participants received an online instruction page. They were asked to perform the following steps: (1) Use a prototype developed for this study (see Figure1) to select a title from a list of 30 YouTube videos or 30 Economist GDs. We restricted the media to 30 to ensure that each media was approximately the same length and quality. The participants then chose and watched a video or read a GD article. (2) Use the prototype to retrieve a ranked list of the top 30 comments for a video or a GD based on reverse-chronological order and vote on each comment. (3) Use the prototype to retrieve a ranked list of the top 30 comments for a video or a news article based on crowd-based order (these were the top ranked comments highlighted by YouTube or recommended by The Economist) and vote on each comment. (4) Use the prototype to retrieve a ranked list of the top 30 comments for the same video in accordance with their preferences by selecting combinations of facets and topics and vote on each comment. Participants were given 5 minutes in each step. In the facet-based ranking, each comment is voted along two dimensions: (1) **Interestingness:** If it contains interesting content for user personally and not necessarily for other users. (2) **Relevance:** If it contains relevant content to user's selected facets and topics. It is important to note that in our setting *relevance* is a meaningless concept without selecting a particular facet. Therefore, in the chronological and crowd-based ranking modes, only the *interestingness* is rated. This is due to the fact that a comment is only considered relevant when it is relevant to the selected facets of the user. For example, when a user selects the subjectivity facet, in our experiment, the relevant comments are comments with higher subjectivity tone, but they are not necessarily relevant to the topic of the video. Additionally, we also explained that the comment does not necessarily have to be directly relevant to the video content.

We restricted the size of the ranked list of comments to 30 in order to minimize judgment fatigue. Additionally, due to the fact that different combinations of facets result in various numbers of comments, we excluded all rankings which resulted in less than three comments (76 rankings out of 462 rankings). For example, if particular combinations for a video result in only one comment which is relevant, then MAP and P@10 are given the value of 1. This result would have skewed our evaluation and for this reason we omitted those instances. Finally, in order to determine the type or types of facets that are most effective, we asked our study participants to explore different combinations of facets:

- **TF**: selecting combinations of only topic facets.
- **SF**: selecting combinations of only subjective facets.
- **OF**: selecting combinations of only objective facets.
- **Any**: any combination of facets.

To measure the effectiveness of the different facets, we rely on standard information retrieval effectiveness measures: *Mean Average Precision* (MAP) as well as *Precision* at 10 and 20 documents respectively (P@$\{10,20\}$). With these measures, we consider the use case that a user is interested in finding many relevant and interesting comments for each facet selection. While P@k is a set-based measure, MAP takes the ranking of relevant/interesting items into account as well.

Results: Figure 2 shows the effectiveness of the faceted ranking compared to the prevalent ranking methods. In total, participants votes on 462 ranking lists (375 for YouTube dataset and 87 for Economist dataset). When considering the first default ranking method,

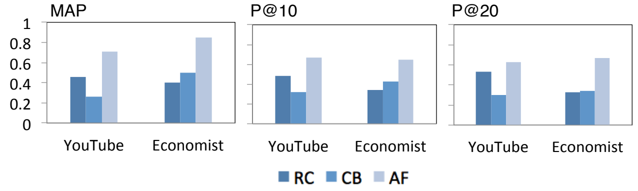


Figure 2: Comparison of the effectiveness of the adaptive faceted (AF) ranking to two prevalent ranking methods: Reverse-Chronological (RC) and Crowd-based (CB)

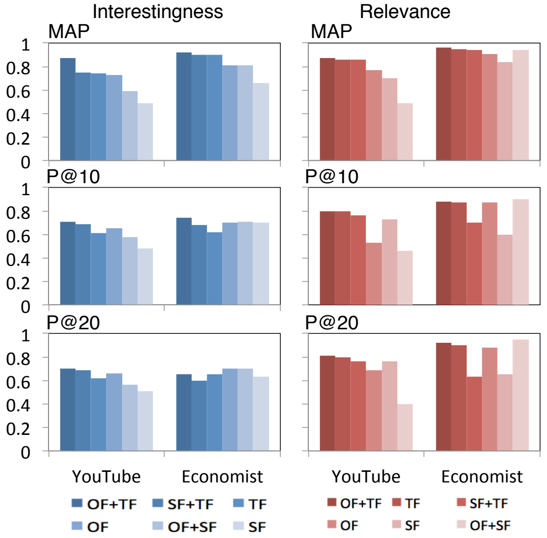


Figure 3: Comparison of effectiveness of different strategies for the selection of different types of facets. Values sorted by MAP values in decreasing order, therefore legends are different.

the reverse-chronological ranking, the measures indicate that this ranking is at least partially effective for both platforms (MAP of .46 and .41 for YouTube and Economist respectively). Approximately half of the comments retrieved are *interesting* to the users. Furthermore, in consideration of the second ranking method, crowd-based ranking, the effectiveness measures indicate that the effectiveness of this ranking type is associated with the culture of the platform. It is less effective compared to the reverse-chronological ranking for the YouTube dataset, while it is more effective than reverse-chronological ranking for the Economist dataset (MAP of .26 and .51 for YouTube and Economist respectively). Approximately one third of the comments retrieved for YouTube set and half of the comments retrieved for Economist set are determined to be *interesting* to the users. In contrast, in the ranking of comments retrieved with our adaptive faceted ranking strategy, approximately every two out of three results are deemed to be interesting for both platforms (MAP of .71 and .85 for YouTube and Economist respectively). Thus, overall we have shown that the adaptive faceted ranking method yields considerable improvements in terms of effectiveness compared to prevalent ranking methods.

Figure 3 shows the effectiveness of combinations of different types of facets. We observe that the combinations which exploit topic and objective facets (OF+TF) considerably outperform other combinations with respect to *interestingness* (MAP of .87 and .92 for YouTube and Economist respectively). However, the other strategy which exploits topic and subjective facets (SF+TF) performs only slightly worse (MAP of .75 and .90 for YouTube and Economist respectively). In particular, we observe that combinations which exploit topic facets (TF, OF+TF, and SF+TF) perform better than combinations without topic facets (SF+OF, SF, and OF), while two combinations which take into consideration subjective facets (SF, OF+SF) perform considerably lower than other combinations. The worst performing strategy is using only subjective facets. This is due to the fact that among faceted rankings based on subjective facets, most of rankings were based on “Offensive and angry” facet. Therefore, most of ranked comments were off-topic and vague and consequently users have less personal interest about these comments. Nevertheless, such a facet can assist the moderators of platforms to filter and extract off-topics and vague comments. The results are similar when considering *relevance* instead of *interestingness*. Also for the *relevance* dimension, we observe that the combination which exploits topic and objective facets (OF+TF) considerably outperform other combinations (MAP of .87 and .96 for YouTube and Economist respectively) and combinations which exploit topic facets perform better than combinations without topic facets.

While these results are useful for an initial analysis, a statistical analysis is required to draw more confident conclusions about how the effectiveness of strategies differs.

For each dataset we considered 3 sets: (1) all positive and negative interesting votes collected for the faceted ranking results, (2) all positive and negative interesting votes collected for the reverse-chronological ranking results, and (3) all positive and negative interesting votes collected for the crowd-based ranking results. Then we applied 2 Pearson's Chi-squared tests (1) between the first and second sets, and (2) between the first and third sets. The results of this study (shown in the top part of Table 3 indicate that adaptive faceted ranking is statistically significantly different from reverse-chronological ranking (RC) and crowd-based (CB) ranking. These results confirm our hypothesis which proposed that faceted ranking would outperform the two other ranking methods. Furthermore, we also present the mean and standard deviation values of positive votes on comments for different facets to help understand the magnitude and direction of the differences.

Additionally, we applied Pearson's Chi-squared tests between sets related to different adaptive faceted rankings with varying combinations of types of facets. More precisely, we applied a test for each set related to each combination of facets and the set related to all types of facets (without taking into consideration the type of facets). The results of this study (shown in the lower part of Table 3) indicate that the adaptive faceted rankings which use a combination of topic-related facets and objective facets are statistically significantly different from the faceted ranking with other combinations of facets. These results confirm that the combination of objective and topic facets is more effective when compared with combinations of other types of facets. Furthermore, it is indicated that the adaptive faceted rankings, which use a combination of only subjective facets, are statistically significantly different from the faceted ranking with

Ranking	YouTube			Economist		
	X2	Mean	STD	X2	Mean	STD
RC	25.26**	12.2	5.7	105.4**	10.15	7.65
CB	90.58**	8	2.2	103.8**	10.38	7.55
AF (OF+TF)	68.07*	21.8	7.8	5.575*	25.9	8.6
AF (SF+TF)	4.591	16.7	8.4	0.046	18.5	7.3
AF (OF+SF)	2.358	12	3.6	0.002	14	5.7
AF (TF)	0.056	14.9	7.8	1.401	19.6	8.5
AF (SF)	49.25*	11	6.6	42.27*	12	4.6
AF (OF)	5.787	15	7.8	0.021	20	7.3
AF	–	14.6	8.0	–	21.33	8.7

Table 3: The results of statistical significance tests. Adaptive faceted ranking strategy (AF) is compared to the reverse-chronological (RC) and crowd-based (CB) ranking (upper part), and to different strategies for selecting facets (lower part). All rows for which there is sufficient evidence ($p < 0.0001$) that the two predicted samples come from different distributions are marked with "" and ($p < 0.01$) are marked with "*". The mean and standard deviation of different ranking strategies are shown on the right.**

all types of different combinations of facets. This result confirms that a combination of only subjective facets is less effective when compared with combinations of other types of facets.

Our results demonstrate that combinations of different facets perform better than faceted rankings which use only one type of facet, such as subjective facets or topics alone. We can also conclude that topic-related facets are very important to a successful ranking strategy, in terms of both relevance and degree of interest. Nevertheless, different combinations of objective or subjective facets with topic facets perform slightly better and are more effective when comments do not explicitly represent a specific topic. In addition, objective facets are more desirable than subjective facets or topics alone.Therefore, we suggest that objective and topic facets be given more attention for adaptive faceted ranking by designers of such a system. Also, users should be made aware of objective and topic facets in higher positions of a facet order.

Although adaptive faceted ranking outperformed crowd-based and reverse-chronological ranking methods, we believe that chronological ranking is still useful for users with particular tasks in mind. Therefore, we suggest that chronological ordering may be designed and developed as one of the facets to be suggested to users. Considering crowd-based ranking, the effectiveness of this method is influenced by the culture of the platform. For platforms with productive and constructive communities and discussions, such as The Economist, this ranking is useful and is worth designing as a type of facet.

4.2 Study 2: Semantic Enrichment Performance

As the results of the first study are influenced by the performance of facets, this study investigates a particular aspect of the framework, that is the evaluation of the performance of the semantic enrichment component.

First part of this study evaluates the performance of the subjective and objective facets and, in addition, describes the process of developing the Informativeness classification model. Second part investigates effectiveness of different topic-labelling approaches for topic facets.

Facet	Full agreement	Moderate agreement
Subjective	*"One of the greatest speeches i've ever eared..GRAZIE Steve"*	*'Diana died, Barry manakee died, Kanga tryon died in the same year as Diana, the driver of the white fiat died,......everyone had a connection with the tampax. I wonder what will happen if Kate crosses him"*
Informative	*"Austria and Hungry was a major ally of Germany. They helped the Germans annihilate the russian army."*	*"No, the allies started this mess, it was their incompetence that led to ww2, if they were not so damn hard on Germany there wouldn't be a mad man like Hitler coming to power"*
Affective	*"Such an awful thing to happen to such a peaceful and talented man :(R.I.P John Lennon."*	*"If there was one thing everyone involved in the war could agree on, it's that they did not like Versailles."*

Table 4: Examples of comments that achieved full vs. moderate annotator agreement. The three facets shown are those with the lowest inter-annotator agreement.

4.2.1 Evaluation of Subjective and Objective Facets: ***Crowdsourced annotated comment*** In order to create a training set and evaluate the accuracy of our semantic enrichment approach, we created a ground truth dataset by annotating a subset of the comments. Specifically, for ten randomly selected media objects, 100 comments were randomly chosen (thus in total $1,000$ comments). The comments were annotated with respect to *Informative*, *Subjectivity*, *Affect*, *Offensiveness*, *Video Timestamp*, *Sadness*, *Offensive and Anger*, and *Religious referenced*. For the annotation process, we relied on crowdsourcing and employed workers via Amazon's Mechanical Turk[3] platform. To ensure worker quality and attention, workers had to answer two objective questions per task ("1-What is the first word of the comment?" and "2-Select 1-4 keywords that represent the most important terms in the comment"). The answers to these questions can be computed automatically. Those who did not answer the quality questions satisfactorily were excluded from further participation. Additionally, workers had to provide binary judgments for each of the seven facets listed above. Thus, overall nine questions (including two honey pot questions) had to be answered by each worker for each comment.

Having collected three judgments per comment, we first determine the inter-rater agreement for each facet based on Fleiss' Kappa. The results are shown on the left side of the Table 5. The agreement is close to perfect for the *Video Timestamp* facet. This is not surprising, considering the unique syntax of a timestamp. The agreement is also high for *Offensive and Angry* comments while workers had most difficulty to agree on *Subjective*, *Informative* and *Affective* comments. Examples of comments labelled with these facets (with high and low agreement) are shown in Table 4. Overall, we consider a Kappa above .65 for all facets. Comments were labeled along different facets based on majority agreement (when two out of three coders agreed).

Training setup of informative classification model: *"Informative"* facet exploits a trained classification model. This part reports how we trained this classifier. For training the informative classifier, we selected a balanced set of 400 informative comments and 400 NOT informative comments from our crowdsourcing-based annotated comments. Only comments for which at least two out of three coders agreed were selected. We investigated the performance of two classifier models, logistic regression and Naive Bayes. Classifiers were trained using combinations of basic features described above (see Table1). Best performance is achieved by using a subset

[3] https://www.mturk.com/mturk/

Facet	Agreement	Precision	Recall	F1
Video Timestamp	0.97	0.91	0.91	0.91
Subjectivity	0.67	0.95	0.98	0.97
Sad	0.78	0.78	0.65	0.71
Religious referenced	0.76	0.58	0.88	0.70
Offensive and Angry	0.79	0.66	0.90	0.76
Affective	0.75	0.92	0.91	0.92
Informative	0.70	0.86	0.61	0.71

Table 5: Overview of enrichment performance across all facets ordered by their accuracy. Second column shows coders' inter-agreement for each proposed facet based on Fleiss' Kappa.

of features ("Readability", "#Named Entities", "Linkage Variety", "Subjectivity Tone", and "Term Conformity") and the logistic regression classifier (.86 precision , .61 recall, and .71 F1). To determine influential features, in addition to interpreting the statistically significant coefficients of the best performing logistic regression model (using all sets of features), we also ranked the best performing features according to their Information Gain Ratio (IGR).

Results: semantic enrichment performance The performance of our semantic enrichment approach is shown in Table 5. We measure the performance of our semantic enrichment approach by Precision, Recall, and F1 for each facet. It is evident that our approach is highly effective for a number of facets. We are able to achieve a high F1 score, coupled with high precision and recall for enrichment related to the facets *Subjectivity*, *Affective*, and *Video timestamp*. However, the enrichment of facets *Informativeness*, *Religious referenced*, *Sad*, *Offensive and angry* proves to be more difficult. This difference in enrichment performance is a reflection of the difficulties that the human annotators experienced with the task.

4.2.2 Evaluation of Topic-based Facets. Topic labels of comments can be used to provide coherent topical facets. However, topics discussed alongside comments are very noisy due to their short nature. Therefore, exploring performance of different topic labeling approaches for selecting the most appropriate approach is useful. In this study, we empirically evaluate three topic labeling approaches:

1. TF-IDF based on Unigrams: The unigrams with the highest TF-IDF score were utilized. This approach does not require external resources and it is not computationally expensive.

2. Entity-Based: The Named Entities (NEs) appearing in a comment are considered to be indicators of the topics the comment discusses. They are ranked in order of their frequency of occurrence. For the extraction of NEs, we employed the semantic enrichment service AlchemyAPI [4].

3. Topic Modeling (LDA): Finally, we experimented with statistical topic modeling, in particular Latent Dirichlet Allocation (LDA) [3]. An LDA model was trained by aggregating all comments on a media object and inferring the topic distribution from this aggregate (the following standard hyper-parameters were used: $\alpha = 50/T$, $\beta = 0.01$ and T = 1000). From the proposed topics produced, the comment was labeled with the term carrying the highest weight.

Crowdsourcing-based evaluation In order to evaluate the three approaches, we randomly picked $1,000$ of our available comments and presented the comments along with their extracted topic labels to Amazon Mechanical Turk workers. For example for a comment: *"For my money Mullen is the Tony Hawk of technical street skating. he pretty much invented it. First on a freestyle board, and then went on to make normal sized boards his bitch."*, three label topics *"Board"*, *"TONY HAWK"*, and *"Matter"* were extracted using TF-IDF based, Named entity based, and LDA based respectively. Or for a comment *"i lie. sometime i speak of the truth. Every lie and truth has a plan and meaning. "* , two label topics *"Lie"*, *"War"* were extracted using TF-IDF based and LDA based, as comment does not contain any named entity. Each worker was shown a comment and a proposed topic label (selected from one of our three approaches). The workers had to answer three questions about the comment — two questions regarding quality (the same quality questions used in our crowdsourcing-based annotating phase in feature-based evaluation to put off MTurk spammers.) and one question regarding the relevance of the topic label to the comment. In this setup, we make use of binary relevance assessments. Thus, for each of the $1,000$ comments we generated three topic recommendations and collected three worker judgments for each pair of topic label and comment.

Results: topic labeling For the purposes of this study, the outcomes are binary. When considering comments with one or more Named Entities (among $1,000$ comments only 420 comments contain Named Entities), the error rate is 3.85% for Entity based, 26.93% for TF-IDF based, and 69.67% LDA based topic labeling. For comments without Named Entities (67% of the comments), the TF-IDF based topic labeling outperforms the LDA based (the error rate is 28.78% for TF-IDF based and is 71.63% for LDA based). We find that the LDA Analysis generally does not provide meaningful topic terms. Also, providing interpretable descriptions for topic models is a difficult problem. These results show that for the problem of extracting a relevant topic label from a comment, the Entity-Based approach performs better than the investigated alternatives for those comments when Named Entities occur. Nevertheless, it should be noted that NER tools are not perfect - extracted Named Entities can be ambiguous and their disambiguation sometimes goes wrong, especially in short texts. This is a limitation for the topic extraction approach based on NEs. Consequently, additional fine tuning is required when relying on Named Entities such as topic proxies.

5 CONCLUSIONS

In this work, using semantic enrichment in a novel adaptive faceted ranking framework results in a set of semantic facets, enabling adaptive ranking of social comments for end-users. Our experimental results demonstrate that adaptive faceted ranking outperforms other prevalent ranking methods and users can easily find interesting comments based on their preferences. Adaptive faceted ranking combining various types of facets outperforms faceted ranking based on a single type of facet. Also, the type of the facet influences the effectiveness of ranking results, that is objective facets are more effective than subjective or topic facets. In the future, we will explore the use of personalized ordering of facets and ranking strategies to further improve the interestingness and relevancy to individual users. This exploration will require further work in user modelling, model similarity, and facet selection optimisation. Finally, different designs and orderings of facets may also have a significant impact on users' interactions with the system and the results achieved, and these are aspects which we will also explore further.

[4] http://www.ibm.com/watson/alchemy-api.html

REFERENCES

[1] Fabian Abel, Ilknur Celik, Geert-Jan Houben, and Patrick Siehndel. 2011. Leveraging the Semantics of Tweets for Adaptive Faceted Search on Twitter. In *Proceedings of the 10th International Conference on The Semantic Web (ISWC'11)*.

[2] Michael S. Bernstein, Bongwon Suh, Lichan Hong, Jilin Chen, Sanjay Kairam, and Ed H. Chi. 2010. Eddi: interactive topic-based browsing of social status streams. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology (UIST '10)*.

[3] D.M. Blei, A.Y. Ng, and M.I. Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning res* (2003).

[4] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web (WWW '11)*. https://doi.org/10.1145/1963405.1963500

[5] Anish Das Sarma, Atish Das Sarma, Sreenivas Gollapudi, and Rina Panigrahy. 2010. Ranking mechanisms in twitter-like forums. In *Proceedings of the third ACM international conference on Web search and data mining (WSDM '10)*.

[6] Liangjie Hong, Ron Bekkerman, Joseph Adler, and Brian D. Davison. 2012. Learning to rank social update streams. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '12)*.

[7] Jessica Hullman, Nicholas Diakopoulos, Elaheh Momeni, and Eytan Adar. 2015. Content, Context, and Critique: Commenting on a Data Visualization Blog. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing (CSCW '15)*. 1170–1175.

[8] Jonathan Koren, Yi Zhang, and Xue Liu. 2008. Personalized Interactive Faceted Search. In *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*.

[9] Cliff A.C. Lampe, Erik Johnston, and Paul Resnick. 2007. Follow the reader: filtering comments on slashdot. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*.

[10] Jingjing Liu, Yunbo Cao, Chin Y. Lin, Yalou Huang, and Ming Zhou. 2007. 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)*.

[11] Elaheh Momeni, Simon Braendle, and Eytan Adar. 2015. *Adaptive Faceted Ranking for Social Media Comments*. Springer International Publishing, Cham, 789–792.

[12] Elaheh Momeni, Claire Cardie, and Nicholas Diakopoulos. 2015. A Survey on Assessment and Ranking Methodologies for User-Generated Content on the Web. *ACM Comput. Surv.* 48, 3, Article 41 (Dec. 2015), 49 pages.

[13] Elaheh Momeni, Claire Cardie, and Myle Ott. 2013. Properties, Prediction, and Prevalence of Useful User-generated Comments for Descriptive Annotation of Social Media Objects. In *The 7th International AAAI Conference on Weblog and Social Media (ICWSM2013)*. AAAI, Boston, USA.

[14] Tim Paek, Michael Gamon, Scott Counts, David Maxwell Chickering, and Aman Dhesi. 2010. Predicting the Importance of Newsfeed Posts and Social Network Friends.. In *AAAI*. AAAI Press.

[15] Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl, and Jose 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, 10. https://doi.org/10.1145/1772690.1772781

[16] Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. (2010). http://homepage.psy.utexas.edu/homepage/students/Tausczik/Yla/index.html

[17] Claudia Wagner, Matthew Rowe, Markus Strohmaier, and Harith Alani. 2012. What Catches Your Attention? An Empirical Study of Attention Patterns in Community Forums. In *ICWSM'12*.

[18] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. 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)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 347–354.