



universität
wien

DISSERTATION

Titel der Dissertation

Adaptive Moderation of User-Generated Content on Web

Verfasserin

Dipl.-Ing. Elaheh Momeni Roochi

Angestrebter akademischer Grad

Doktorin der technischen Wissenschaften (Dr. techn.)

Wien, 2014

Studienkennzahl lt. Studienblatt: A 786 880

Dissertationsgebiet lt. Studienblatt: Informatik

Betreuerin / Betreuer: Univ.-Prof. Dipl.-Ing. Dr. Wolfgang Klas

Abstract

User-generated content on the Web, and particularly in social media platforms, facilitates the augmentation of additional information with digital resources and delivers valuable information. However, some user-generated content (UGC) is not useful due to the varying intentions of authors of content and perspectives of viewers. This raises the challenge of how to maximize its value for platform viewers. Currently, the majority of available approaches tends to train an assessment and ranking approach for maximizing various values such as usefulness, relevancy, or credibility for a platform’s viewers. However, most of these approaches rely on particular sources of ground truth and do not enable moderation requesters to make adaptive assessments of a particular value. Accordingly, there is insufficient consideration of approaches which are adaptive for individual users. Many of the available approaches do not enable individual requesters to adapt a moderation to their requirements. In the attempt to overcome this challenge, this thesis aims to provide researchers and Web data curators with a comprehensive understanding of existing work, thereby encouraging further experimentation and development of new approaches focused on automated moderation of user-generated content. Accordingly, an adaptive moderation framework is proposed. It is a semi-supervised approach which semantically enriches and clusters content along multiple explicit semantic facets (e.g., subjectivity, informative, and topics) and enables users to explore different facets and select combinations of facets in order to extract and rank content that matches their interests.

The development of this framework is the result of the following investigations. First, a systematic review of approaches for assessing and ranking of UGC has been conducted, producing results which have been obtained by gathering and comparing existing approaches. These are grouped in three categories: Community-based assessment and ranking of UGC, Single-user assessment and ranking of UGC, and Incentivizing high-quality contributions. Second, in order to provide automated support for the curation of useful user-generated comments when there is no explicit or

implicit feedback from a user, a crowd-sourced gold standard of USEFUL and NON-USEFUL comments has been constructed. Then, standard machine learning methods have been used to develop a “usefulness” classifier, exploring the impact of surface-level, syntactic, semantic, and topic-based features in addition to extra-linguistic attributes of the author and her social media activity. Third, an existing model of prevalence detection has been adapted, using the learned classifier to investigate patterns in the commenting culture of two popular social media platforms. Fourth, a prototype of a Web-based interface implementation of the proposed adaptive moderation framework has been developed, enabling the evaluation of the proposed framework and exploration of different ranking strategies.

The systematic review of approaches for assessing and ranking of UGC has revealed a number of influential text-based and contextual features related to different entities — authors, online social media resources, and content — of social media platforms. These features are shown to be effective for many machine-based methods of assessment and ranking of UGC and motivate a selection of a list of facets for the proposed adaptive faceted moderation framework. The results of the study conducted on the estimation of the prevalence of useful UGC has indicated that the prevalence of USEFUL content is platform-specific and is also influenced by the entity type of the media object being commented on (person, place, event), its time period (e.g., year of an event), and the degree of polarization among content generators. Finally, the results of the evaluation of the proposed adaptive moderation framework show that an adaptive faceted ranking performs significantly better than reverse-chronological ranking and has substantial benefits. These include clustering each element of a comment along multiple explicit semantic facets rather than in a single topic or subjective facets.

Zusammenfassung

Benutzer-generierte Inhalte im Web (“user-generated content” bzw. “UGC”) und im Speziellen in Social Media Plattformen, erleichtert die Erhöhung zusätzlicher Information mit digitalen Ressourcen und liefern wertvolle Informationen. Teilweise ist UGC jedoch aufgrund der unterschiedlichen Intentionen der Autoren sowie der Perspektiven der Leser nicht nützlich. Darauf basiert die Herausforderung, den Wert von UGC für Plattformnutzer zu maximieren. Aktuelle Methoden tendieren dazu, einen Bewertungs- und Rankingansatz zu trainieren, um unterschiedliche Werte wie Nützlichkeit, Relevanz oder Glaubwürdigkeit für die Nutzer einer Plattform zu maximieren. Die meisten dieser Ansätze verlassen sich jedoch auf eine bestimmte “Ground Truth” und ermöglichen im Falle von Moderationsanfragen keine adaptive Beurteilung des bestimmten Wertes. Dementsprechend werden für individuelle Nutzer adaptive Ansätze nicht ausreichend berücksichtigt. Viele der verfügbaren Ansätze ermöglichen es individuellen Bedarfsträgern nicht, eine Moderation an ihre Bedürfnisse anzupassen. Um dieses Problem zu lösen, fokussiert diese Dissertation darauf, Wissenschaftlern und Web Daten Kuratoren ein umfassendes Verständnis bestehender Arbeiten zu vermitteln und dabei weitere Experimente und Entwicklungen neuer Ansätze automatisierter Moderation von nutzergenerierten Inhalten anzuregen. Dementsprechend wird ein Rahmenwerk adaptiver Moderation präsentiert. Es handelt sich um einen semi-überwachten Ansatz, der Inhalte semantisch anreichert und anhand von multiplen, expliziten Aspekten kategorisiert (z.B. Subjektivität, Informationsgehalt sowie Art des Themas) und der es Nutzern ermöglicht, verschiedene Aspekte zu erforschen und Kombinationen von Aspekten zu wählen, um Inhalte, die ihren Interessen entsprechen zu extrahieren und zu ranken.

Die Entwicklung dieses Rahmenwerks ist das Ergebnis der nachfolgenden Untersuchungen. Zunächst wurde ein systematischer Überblick über existierende Ansätze zu Bewertung und Ranking von UGC erstellt. Die Ansätze wurden verglichen und in drei Kategorien zusammengefasst: “Community-based assessment und ranking” von UGC, “Single-user assessment und ranking” von UGC sowie “Incentivizing

high-quality contributions”. Zur automatisierten Unterstützung der Moderation von nutzergenerierten Inhalten bei Fehlen von explizitem oder implizitem Nutzerfeedback wurde in einem nächsten Schritt ein “crowd-sourced gold standard” für nützliche und nicht nützliche Kommentare erstellt. Schliesslich wurden Standard Machine Learning Methoden für die Entwicklung eines Nützlichkeitsklassifizierers, die zusätzlich zu den extralinguistischen Attributen des Autors und seiner “Social Media” Aktivitäten den Einfluss von basistextlichen, syntaktischen, semantischen und themenbasierten Eigenschaften untersucht, herangezogen. Drittens wurde ein existierendes Modell zur “prevalence detection” (Verbreitungsermittlung) adaptiert, das den erlernten Klassifizierer zur Untersuchung von Mustern in der Kommentierungskultur von zwei populären Social Media Plattformen nutzt. Zuletzt wurde ein Prototyp einer web-basierten Schnittstellenimplementierung für das präsentierte adaptive Rahmenwerk entwickelt, wodurch die Evaluierung des präsentierten Rahmenwerks und die Erforschung verschiedener Rankingstrategien ermöglicht werden.

Der systematische Überblick über existierende Ansätze zu Bewertung und Ranking von UGC hat eine Anzahl von Einflüssen text- und kontextbasierter Eigenschaften in Bezug auf unterschiedliche Entitäten — Autoren, Online Social Media Quellen und Inhalte — aufgezeigt. Diese Eigenschaften haben sich für viele machine-based Methoden zu Bewertung und Ranking von UGC als wirksam erwiesen und regen eine Reihe von Aspekten für das präsentierte adaptive facettierte Moderationsrahmenwerk an. Die Ergebnisse der zur Einschätzung der Verbreitung nützlicher UGC durchgeführten Studie haben gezeigt, dass die Verbreitung von nützlichen Kommentaren plattformspezifisch ist und auch durch den Entitätentyp des kommentierten Medienobjekts (Person, Ort, Ereignis), durch die Zeit (z.B. das Jahr eines Ereignisses) sowie durch den Grad der Polarisierung unter den inhaltsgenerierenden Nutzern beeinflusst wird. Letztendlich die Ergebnisse der Evaluierung des präsentierten adaptiven Moderationsrahmenwerks zeigen, dass ein adaptives facettierte Ranking signifikant besser funktioniert als ein reverse-chronological Ranking und substantielle Benefits aufweist. Diese umfassen jedes Element eines Kommentars entlang multipler expliziter semantischer Facetten anstatt in singulären Themen oder subjektiven Facetten.

Acknowledgements

Writing these acknowledgements has taken longer than I had expected and has not been so easy, there being so many people to thank for their support and help during my research. These people shared ideas and gave input, thus shaping and influencing my work. Accordingly, I wish to thank all of them, although it is not possible to actually mention everybody.

First and foremost, I would like to express my deepest appreciation to my advisor Prof. Wolfgang Klas who constantly inspired, encouraged, and advised me while always allowing me the freedom to pursue my research. I truly benefited from his wealth of experience.

Special thanks also go to my PhD committee members. In particular, I would like to thank Prof. Wolfgang Nejdl for his constructive comments and suggestions which contributed to making this thesis more complete and accurate. I would also like to thank Prof. Gerald Quirchmayr for his helpful support during the completion of this thesis.

I also want to sincerely thank Prof. Claire Cardie, Prof. Geert-Jan Houben, and Prof. Eytan Adar. I feel truly blessed that I was fortunate enough to be accepted as a research visitor at their research groups and work with such intellectually inspiring researchers. Discussions with them gave me the opportunity to develop a new perspective on my work and I learned so much in such a short period of time. Also, thank you for making me feel very welcome right from the start. I also want to express my gratitude to the Natural Language Processing group of the Cornell University, the Web Information Systems group of the Delft University of Technology, and the Michigan Interactive & Social Computing group of the University of Michigan. During the times spent with all these groups, not only did I proceed with my research but I also made new friends and enjoyed my visits.

Furthermore, I would like to thank all my university and research colleagues with whom I authored a number of papers, especially Dr. Bernhard Haslhofer, Dr. Myle

Ott, Ke Tao, Dr. Claudia Hauff, and Gerhard Sageder for many discussions, valuable feedback and fruitful collaboration. Also, I appreciate the help and support given to me by Dr. Reza Rawassizadeh, Dr. Michael Sedlmair, and Michael Flynn. Each of them helped me and I have benefited from their knowledge. And, of course, I would also like to thank the students who have contributed to this work via their projects and theses. In particular, I would like to thank Simon Braendle. I really enjoyed working with him and supervising his thesis.

Besides, I would like to express my thanks to current and former members of the Multimedia Information Systems group of the University of Vienna for their help and assistance. Especially, I want to thank Jan Stankovsky, Peter Kindermann, and Manuela Schena for their support.

Furthermore, I thank my husband, Matthias Ortner, from the bottom of my heart. Without his constant love, support and presence, this thesis would not have been possible. Words are not enough to express how grateful I am.

Last, but not least, I wish to express my sincere gratitude to my beloved parents and sister, Parinaz, who supported me throughout my years of study and in every aspect of life, and who encouraged me greatly during this thesis.

Contents

1	Introduction	2
1.1	Background and Problem Description	2
1.2	Thesis Goal	6
1.3	Organization of the Thesis	11
2	State-of-the-art-analysis	15
2.1	Introduction	15
2.2	Notion of value, expected to be maximized by assessment and ranking approaches	18
2.3	Influential Features Taxonomy	22
2.4	Community-Based Assessment and Ranking of UGC	25
2.4.1	Crowd-based method	25
2.4.2	Machine-based method	27
2.4.3	Summary	47
2.5	Single-User Assessment and Ranking of UGC	51
2.5.1	Personalized Approaches	52
2.5.2	Interactive & Adaptive Approaches	53
2.5.3	Summary	54
2.6	Incentivizing high-quality User-generated Content	56
2.7	What Do We Observe, and Where Do We Need Deeper Focus	57

3	Experiments and Datasets	63
3.1	Introduction	63
3.2	Features Engineering	67
3.3	Data Acquisition	75
3.3.1	List of Topics	76
3.3.2	Datasets	76
3.3.3	Collecting User Judgements for Defining Usefulness	77
3.3.4	Collecting Expert Judgements for Defining Usefulness	80
3.4	Experiments	82
3.4.1	Usefulness Classifier	83
3.4.2	Influence of Features on Usefulness Classifier	84
3.4.3	Influence of Topic on Classification	87
3.4.4	Influence of Commenting Culture of Platforms on Characteristics of Useful Comments	90
3.4.5	Prevalence of Useful Comments	92
3.5	Discussion	99
4	Requirements and Design	102
4.1	Introduction	102
4.2	Design Considerations	103
4.2.1	Fundamental Design Aspects	103
4.2.2	Conclusions for Design Decisions	105
4.3	AMOWA: A Framework for Adaptive Moderation of UGC	107
4.3.1	Component 1: Semantic Enrichment	107
4.3.2	Component 2: Facet Extraction and Ranking	109
4.3.3	Component 3: Feedback Collector and Optimization	111
4.3.4	Component 4: Baseline Usefulness Model	113
4.4	Functional Specification of AMOWA by Means of Services	114
4.4.1	Interface Specification — SEFE Service	114

4.4.2	Interface Specification — Ranking Service	117
4.4.3	Interface Specification — Feedback Service	117
4.5	Summary	118
5	Implementation	120
5.1	Introduction	120
5.2	Usefulness Prediction Model	121
5.3	AMOWA-WS (A Web Service for AMOWA)	122
5.4	AMOWA-UI (A User Interface for AMOWA)	130
5.5	Summary	133
6	Experiments and Evaluation of Proposed Framework	136
6.1	Introduction	136
6.2	Study 1: Effectiveness of Adaptive Faceted Ranking Strategies	138
6.2.1	Participants	138
6.2.2	Experimental Setup	139
6.2.3	Results of Quantitative Assessment	140
6.2.4	Results of Subjective Assessment	144
6.3	Study 2: Faceted Extraction and Ranking Performance	147
6.3.1	Experimental Setup	147
6.3.2	Results	148
6.4	Study 3: Comparison of Topic Detection Algorithms	149
6.4.1	Algorithms	150
6.4.2	Experimental Setup	151
6.4.3	Results	152
6.5	Discussion	153
7	Conclusions	155
7.1	Discussion and Experimental Results	155

7.2	Conclusions and Future Directions	165
7.2.1	Limitation and Future Work	165
7.2.2	Summary and Conclusions	166
8	Appendices	169
8.1	Appendix1 – Experimental Datasets of Related Work	169
8.2	Appendix2 – Online Evaluation Instruction	175

List of Figures

1.1	Enriched Photo by UGC	3
1.2	Examples and comparison of ranking methods for UGC	10
2.1	Evolution of approaches related to the assessment and ranking of UGC	19
2.2	Values which are important and assessed by different application do- mains.	22
2.3	Features Taxonomy	23
2.4	Videos with low versus high variance of comment ratings.	31
2.5	Helpfulness voting system of Amazon online shopping service	35
2.6	Overview of community-based assessment and ranking of UGC ap- proaches.	48
2.7	Influential features sets for assessment and ranking of different values	50
2.8	Overview of single-user assessment and ranking of UGC approaches .	55
2.9	Overview of ranking and assessment approaches of UGC.	58
3.1	Examples of photos of the Library of Congress on Flickr Commons .	65
3.2	Performance results of classification using top-20 features	88
3.3	Graph of Bayesian estimates of usefulness prevalence versus time pe- riods.	95
3.4	Graph of Bayesian estimates of usefulness prevalence versus polariza- tion of topics.	97
3.5	Different platforms and topics lead to different usefulness prevalence.	98

4.1	Abstract overview of proposed adaptive moderation framework. . . .	108
4.2	Interface specification of SEFE service	115
4.3	Interface specification of Ranking service	116
4.4	Interface specification of Feedback service	118
5.1	Screenshot 1 of the Web-based user interface of the framework	131
5.2	Screenshot 2 of the Web-based user interface of the framework	132
5.3	Screenshot 3 of the Web-based user interface of the framework	132
5.4	Ranked comments by selecting a combination of objective and topic facets.	134
5.5	Ranked comments by selecting a combination of subjective and topic facets.	135
5.6	Ranked comments by selecting a combination of topic facets	135
6.1	Percentages of comments with various combination of “Interesting” and “Relevant” votes.	144
6.2	Overview of the subjective evaluation	145

List of Tables

2.1	Sample of tweets with high retweets.	33
2.2	Sample of deceptive and truthful reviews.	39
2.3	Samples of Request and Introduction posts	43
3.1	Overview of Features	67
3.2	Summary statistics of dataset crawled from Flickr Commons	77
3.3	Summary statistics for datasets	77
3.4	Manual coding results across platforms.	78
3.5	The comparison of the mean and standard deviation values of each feature between useful (U) and non-useful (N) comments.	79
3.6	The comparison of the mean and standard deviation values of each feature between user-judged (U) and expert-judged (E) useful comments.	81
3.7	Evaluation of classification algorithms	84
3.8	Coefficient ranks of top-20 features	85
3.9	Results from the evaluation of usefulness classifiers for different object types.	89
3.10	Coefficient ranks of top-20 features for each topic.	91
3.11	Significant differences between prevalence of usefulness for various topics related to different time periods.	96
3.12	Significant differences between prevalence of usefulness for various topics with various polarization values.	96

3.13	Significant differences between prevalence of usefulness for various topics with various polarization values.	99
4.1	Overview of proposed facets.	110
5.1	Description and arguments of the Web-based interface related to SEFE service	123
5.2	Description and arguments of the Web-based interface related to the Ranking service	125
5.3	Description and arguments of the Web-based interface related to the Feedback service	126
5.4	Description and arguments of Web-based interface related to Login service	127
6.1	Basic statistics of experimental data set (YouTube videos and comments).	138
6.2	Effectiveness of adaptive faceted ranking strategies	141
6.3	Performance of adaptive faceted ranking	143
6.4	Judges' inter-agreement for each proposed facet based on Fleiss' Kappa. 148	
6.5	Examples of comments that achieved full vs. moderate annotator agreement.	149
6.6	Overview of clustering performance across all facets ordered by their accuracy.	150
6.7	Topic label examples. Bolded items shows topic labels with highest agreement among coders	151
6.8	Coefficients and Odds-Ratios of different topic labeling approaches evaluated on 1000 comments.	153
8.1	Short overview of main contributions and experimental datasets of related work.	169

Chapter 1

Introduction

1.1 Background and Problem Description

User-generated Content (UGC) on the Web and in particular on social media platforms is a vital part of the online social media ecosystem [Asselin et al., 2011, Rotman et al., 2009, Rangwala and Jamali, 2010]. UGC provides a way for participants to “evolve” multimedia social objects — ranging from YouTube videos to News articles — by contributing multiple perspectives and observations, answering questions, forming hypotheses, and otherwise contributing to the development of the “social object” [Shamma et al., 2007]. This helps to find new trends and discover knowledge about the end-users who generate content. It can also facilitate machine-based processes such as recommendation, retrieval, and search processes.

In the context of multimedia information systems, *descriptive annotations* for social media objects (such as an online video or photo) by experts provide important supplemental information about an object (e.g., textual documents, images, videos) in the form of keywords and free-form descriptions [Golder and Huberman, 2006, Halpin et al., 2007]. Usually comprehensive and of high quality, expert annotations are valuable both for human consumption, aiding efficient information retrieval, and resource management. However, they are costly to create. UGC, on the other hand, represents a potential complementary source of essential information like the

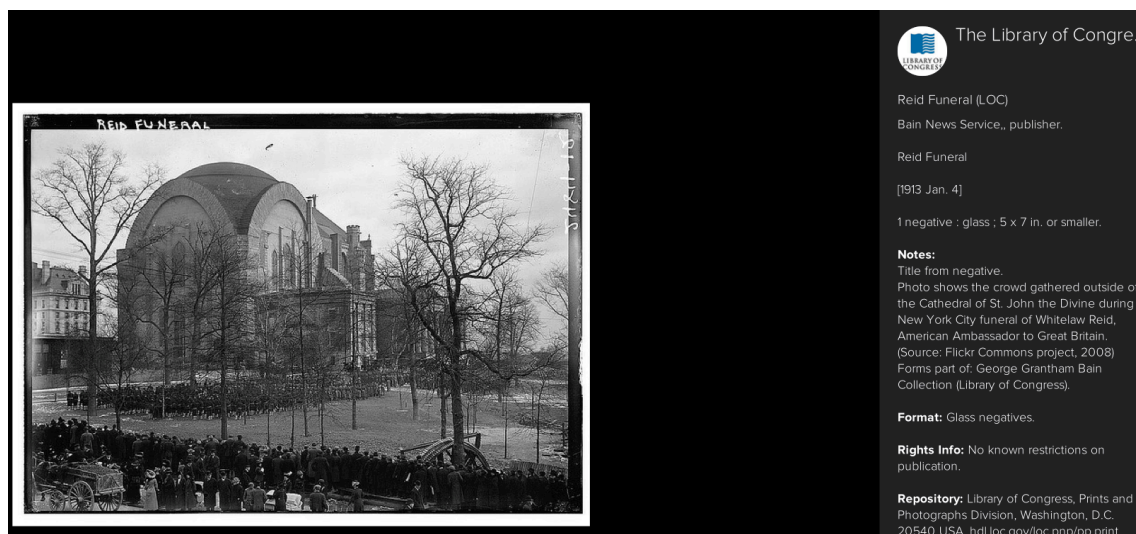


Figure 1.1: Flickr Commons photo - Reid Funeral (a photo of the Library of Congress collection). Description of photo is fully enriched by the user-generated comment

names and places depicted in a photo or video — information that is often not available in existing metadata records [Ames and Naaman, 2007, Kennedy et al., 2007, Asselin et al., 2011, Rotman et al., 2009]. For example, Flickr Commons allows libraries and museums to share their resources so that users can collaborate in the creation of descriptive annotations. One example of the results of this project is a photo from a set of the Library of Congress on Flickr shown in Figure 1.1 that was originally captioned simply as “Reid Funeral”. It is now more fully described by the user-generated comment¹:

Flickr Commons photo - Reid Funeral (LOC)². “Photo shows the crowd gathered outside of the Cathedral of St. John the Divine during New York City funeral of Whitewall Reid, American Ambassador to Great Britain.”

Unfortunately, most UGC presentation systems are simple temporal streams that contain a diversity of focus, usefulness, and quality. Users have different backgrounds, levels of expertise, and intentions for contributing content. As a result,

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

²http://www.flickr.com/photos/library_of_congress/2515741281/

the quality of user-generated content varies from very useful to entirely useless, and UGC can even be abusive or off-topic. Managing and hosting this content can be costly and worse, due to the substantial amount of content, moderation of content is often both time-consuming and challenging. Without a mechanism for end-users to disentangle content streams and identify those likely to be of interest, it is easy to imagine most users being overwhelmed by and disappointed with their experience and worse, to stop participating themselves. Therefore, the task of automatic moderation of UGC to maximize value for the platform’s viewers is becoming increasingly important.

Moderation of UGC is a relatively complex task due to the fact that:

1. UGC is a relatively general term which can refer to different application domains such as tags, product reviews, postings in the questions & answers platforms (Q&A) and discussion forums, comments on digital resources, and so on. Each type of user-generated content has different characteristics.
2. What is defined as value varies with regard to different characteristics of application domains and specific tasks in hand. For example, extracting relevant tweets related to a specific news topic is an important value in micro-blogging platforms whereas extracting truthful product reviews is an important value in product reviews.
3. A particular value can be assessed and maximized in different ways due to different characteristics of UGC. For example, assessing credibility of product reviews requires different features and methods compared to extracting credible postings in micro-blogging platforms. Product reviews can be long and deceptive which are written so that the reader believes they contain the truth, but instead they actually give the reader false information. Therefore, the features related to the text of a review are important features to assess review credibility [Ott et al., 2011]. Instead, postings in micro-bloggings might be short and features related to texts on their own can not help to assess the credibility of postings [Castillo et al., 2011, Morris et al., 2012]. So, we require features to be included which relate to user activities and backgrounds for more accurate assessment. Also, moderation can depend on a number of

factors including the media type (e.g., document, video, art object, photo), the entity type of the object (e.g., the object is associated with a person, place, or event), the time period associated with the object (e.g., early 20th century vs. the 1960's), or even the degree of controversy surrounding the object.

In spite of these complexities, methods for moderation of user-generated content are gaining increasing attention [Siersdorfer et al., 2010, Diakopoulos et al., 2012, Momeni et al., 2013a, Momeni, 2010, Momeni, 2012]. The simplest method to provide moderation is simply to ask end-users [Siersdorfer et al., 2010, Hsu et al., 2009, Lampe and Resnick, 2004]. This wisdom-of-the-crowd approach simply allows all users to vote on (thumbs up or down, stars, etc.) or rank comments. This avoids an explicit definition of usefulness. Additionally, Liu et al. [Liu et al., 2007] show that voting is influenced by a number of factors (e.g., a “rich get richer” phenomenon) that may distort accuracy.

An alternative method for moderation of user-generated content takes into consideration an explicit definition of a specific value using a machine-based approach such as supervised or unsupervised learning. However, most of the available approaches rely on particular sources of ground truth and do not enable moderation requesters to make personal assessments of a particular value. In other words, there is less consideration of a personalized definition of the value for an individual user and many of the available approaches do not enable individual requesters to adapt the moderation to their requirements. For example, most of the work on identification of helpfulness of product reviews (as a value which is expected to be maximized) creates and develops prediction models based on a set of majority-agreement labeled reviews. However, helpfulness is a subjective concept that can vary for different individual requesters. Therefore, it is important that systems help individuals to make personal assessments of a particular value and provide adaptive moderation which uses different methods to accommodate the differences between individuals with regard to individual requirements and interests. A system should help individuals adapt the moderation based on the particular objective currently in the user's mind.

Therefore, the general challenge we face in this thesis is how to automate moderation

of UGC with regard to the particular objective currently in the user’s mind or the user’s preferences. Our general challenge manifests itself in a number of main research challenges, such as:

1. What are the values expected to be maximized in the moderation process for different application domains?
2. Which moderation methods are effective for maximizing a particular value of UGC with regard to an application domain and user’s preferences?
3. How does moderation adapt to user’s preferences or an objective in the user’s mind?
4. How can we take advantage of semi-supervised learning such as active learning for efficient integration of the crowd into machine-based approaches, or how can we utilize the crowd to optimize the process of moderation and improve the accuracy of hard machine-based judgments?
5. Which features and metrics of the platform are most adequate for moderation of a particular value of UGC with regard to user’s preferences?
6. How well does adaptive moderation, which operates based on user’s preferences, compare to the most prevalent default UGC ranking methods (such as reverse-chronological)?

These main research challenges raise a number of other detailed challenges, which are dealt with in different chapters of this thesis.

1.2 Thesis Goal

The main goal of this work is to provide alternative, *automated* support for the *multi-faceted adaptive moderation* of UGC on the Web. The proposed approach, which is influenced by past work on multi-faceted search [Koren et al., 2008], active learning, and topic identification, is a semi-supervised learning approach for adaptive

moderation of social media content with regard to the preferences of each individual user. We build our adaptive moderation framework on the requirements derived from an analysis of current approaches in assessment and ranking methods of user-generated content. From this framework, we derive further artifacts, namely, a concrete application programming interface and a concrete representation of the prototypical Web-based user interface. In addition, we aim to better understand the characteristics of useful user-generated content and to estimate their prevalence across social media platforms.

In this thesis, we apply the design-science research method [Hevner et al., 2004]: The design-science paradigm “seeks to extend the boundaries of human and organizational capabilities by creating new and innovative artifacts” [Hevner et al., 2004]. IT artifacts are broadly defined as models (abstractions and representations), methods (algorithms and practices), and instantiations (implemented and prototype systems).

Our core contribution is a semi-supervised learning approach that bridges the conceptual gap between an individual moderation requester and machine computation via a so-called adaptive moderation. From a high-level perspective, our intended approach is based on users interacting with the AMOWA (Adaptive Moderation of Web Annotation) framework via a Web browser. The proposed moderation framework clusters each individual piece of content (such as comments on an online media object) along multiple explicit semantic facets (e.g., subjective comments, informative comments, and topics), selects sub-set of facets, and enables end-users to explore facets and rank content with regard to their preferences and interests. This enables the clustering to be accessed and ordered in multiple ways rather than in a single topic order [Bernstein et al., 2010, Abel et al., 2011]. It also avoids having to rely on particular majority-agreement sources of ground truth. The baseline component of the framework is a usefulness prediction model [Momeni et al., 2013a] which is trained based on majority agreement of users for useful content (The system uses this model as the baseline if the user does not explicitly or implicitly give the system feedback). Starting from a possibly empty set of manually labelled content, an algorithm provides clusters of content and, accordingly, proposes relevant facets. Users explore different clusters (different facets such as topics discussed among comments,

subjective opinions, etc) and select combinations of facts in order to filter and extract content that matches their interests. Furthermore, the framework provides users with the possibility to explicitly give feedback and provide a label for each comment. Positively and negatively labeled comments, which are accumulated, are used by the system for improving the clustering model, facet selection, and modeling preferences of a user.

To enable the realization of this framework, first we develop the usefulness predictor model, which is trained to identify useful content based on the majority-agreement of users and used as the baseline component of the framework. Second, we develop a novel technique for clustering content along various semantic facets, which enables multi-faceted moderation of content. Furthermore, in order to provide coherent clusters of content emerging from discussions about topics as potentially useful facets, state-of-the-art topic identification methods are experimented with in order to find the most accurate one based on our use-case.

Given our overall approach and the challenges we face in the context of multimedia object sharing platforms (such as Youtube, Flickr), we can identify the following contributions of this work and previous related publications:

- ***C1: We carry out a comprehensive state-of-the-art analysis of the existing methods and approaches for assessment and ranking of UGC.*** The scope of considered UGC comprises user-generated content short texts (such as product reviews, tags and comments on online multimedia resources, Tweets, etc.) in different application domains (such as product reviews, Micro-Blogging, comments on online media objects, online questions and answers, etc).
- ***C2: We gather a dataset of comments on online multimedia objects*** from popular social media platforms (Flickr and Youtube) and collect users' and experts' usefulness judgments (by using a crowd-sourcing approach) in order to identify the usefulness and various semantic dimensions (such as Subjective, Affective, Offensive, etc) of gathered comments.
- ***C3: We conduct different experiments for identification of the***

characteristics of useful comments. First, we identify technical features that can be derived from textual content and the author’s context and then characterize the usefulness of a comment. Second, we apply the technical features in a series of experiments to build a classifier model that can automatically identify the usefulness of comments. Third, we investigate to what extent certain topics of media objects play a role with regard to the “usefulness” classification.

- *C4: We draw a number of conclusions and requirements for an adaptive moderation framework* which we call the AMOWA (Adaptive Moderation of Web Annotations) framework, from the state-of-the-art analysis and experiments we have conducted. This framework is capable of representing a wide range of requirements for adaptive moderation of UGC and we present this framework on an abstract and conceptual level. Furthermore, we propose a number of strategies for extracting novel facets and topics from social media comments that operationalize the complex dimensions of usefulness.
- *C5: We further anticipate implementations of the proposed framework* by building a solid basis for implementations of our framework, specifying a generic application programming interface that covers static and dynamic aspects of our proposed framework. This specification allows for the implementation of the envisioned moderation framework in a number of application domains.
- *C6: We develop a Web-based implementation of proposed framework, AMOWA (Adaptive Moderation of Web Annotations) framework* that allows users to work with the moderation framework using interaction metaphors. This interface enables end-users to moderate social media content based on their preferences and interests. Users provide feedback simultaneously by implicit means (using the faceted browser) or explicit means (voting). Figure 1.2-(a) shows a list of comments on a YouTube video³ which are ranked based on the default ranking setting of the system (reverse-chronological ranking). On the other hand, Figure 1.2-(b) shows a list of

³https://www.youtube.com/all_comments?v=UF8uR6Z6KLc



Figure 1.2: Examples and comparison of ranking methods for UGC. Part (a) shows the default ranking method used by YouTube (reverse-chronological ranking) and part (b) shows multi-faceted adaptive ranking method proposed by our framework. The proposed framework semantically enriches each comment, clusters the comments, and finally presents a list of facets on the left side of the interface. This enables users to select combinations of proposed facets for presenting a ranked list of comments based on the chosen facets on the right side of the interface.

ranked comments on the same video using the proposed framework (by selecting a combination of facets proposed by the framework).

- ***C7: We demonstrate the benefits of the proposed adaptive moderation approach*** for providing end-users with access to useful content through quantitative and qualitative evaluation of the framework.

1.3 Organization of the Thesis

Chapter 2 aims to investigate the varying notions of “value” across different types of UGC by presenting a unifying scheme that includes the commonly used definitions of value in existing research. This chapter mainly deals with and presents the first contribution of the thesis. This is achieved by answering the following general research questions: What are the values expected to be maximized for different application domains? How are they defined with regard to the particular application domain and the task in hand? What methods are used for assessing and ranking UGC? Which methods are effective for maximizing the value of UGC with regard to an application domain? What are the effective features and metrics used to predict and measure the particular value of UGC? In order to answer these questions, the findings of a systematic review of existing approaches and methodologies for assessing and ranking UGC are presented. The focus is, in particular, on the short, text-based user-generated content typically found on the Web. The discussion and results of this chapter are under review for a journal article [Momeni and Cardie, 2014] (to be titled “A Survey on Assessment and Ranking Methodologies for User-Generated Content on Web”) and were partially published as an article [Momeni, 2012] (entitled “Semi-automatic Semantic Moderation of Web Annotations”).

Chapter 3 gives an overview of different experiments carried out to identify the characteristics of useful comments and create usefulness models. This chapter deals with and presents second and third described contributions. The goal of the work reported in this section is to provide *automated* support for the curation of useful user-generated comments for use as descriptive annotations for digital media objects. In order to achieve this, this chapter investigates four contributions:

1. *Identification of the characteristics of useful comments:* we study two types of media objects (images and videos) from two popular social media platforms (Flickr Commons and YouTube respectively) and collect users' and experts' usefulness judgements (by using a crowd-sourcing approach) to identify the usefulness of crawled comments. We then identify technical features that can be derived from textual content and the author's context, and characterize the usefulness of a comment.
2. *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.
3. *Study the correlation between the commenting culture of a platform with usefulness prediction.* We investigate to what extent the commenting culture of a platform plays a role with regard to usefulness classification.
4. *Study important factors for estimating the prevalence of useful comments.* We adapt an existing model of prevalence detection [Ott et al., 2012] that uses the learned usefulness classifier to investigate patterns in the commenting culture across social media platforms and different dimensions (entity type, time period, and polarization) of topics of media objects.

The discussion and results of chapter 3 were published in several conferences and in a journal article: [Momeni et al., 2013a] (entitled “Properties, Prediction, and Prevalence of Useful User-generated Comments for Descriptive Annotation of Social Media Objects”), [Momeni et al., 2013b] (entitled “Identification of Useful User Comments in Social Media: A Case Study on Flickr Commons”), a journal article [Momeni et al., 2014b] (entitled “Sifting Useful Comments from Flickr Commons and YouTube”) and a short paper [Momeni and Sageder, 2013] (entitled “An Empirical Analysis of Characteristics of Useful Comments in Social Media”).

Based on our observations and findings from Chapter 2 and 3, Chapter 4 discusses our novel adaptive moderation framework by describing a number of design consid-

erations and requirements. We introduce the basic concepts that we include in the framework and then give a conceptual specification of the framework elements by explaining the architecture of the system and the development of the programming interface of the proposed framework. The discussions of chapter 4 deals with and presents fourth and fifth contributions of this thesis and were partially published as a journal article [Momeni et al., 2014b] (entitled “Sifting Useful Comments from Flickr Commons and YouTube”) and are under review for a publication [Momeni et al., 2014a] (to be titled “Multi-faceted Adaptive Ranking of Social Media Comments”).

After having presented the proposed framework and its elements in various levels of abstraction, Chapter 5 discusses prototypical implementations of the most important parts of the proposed framework. The aims of this chapter are to show the flexibility of the proposed framework and its applicability to different social media platforms. Chapter 5 deals with and presents sixth described contribution. In this chapter, we also outline the architecture and important implementation aspects of each of three prototypes:

1. *Baseline Usefulness Model*: we discuss a prototypical implementation of the baseline model for automatically predicting usefulness of UGC without receiving explicit or implicit users’ feedback.
2. *AMOWA-WS*: we discuss a prototypical implementation of a Web service of a proposed approach which can be simply integrated as a plugin into any social media platform or any platform which deals with UGC. It enables end-users to moderate content with regard to their personal interest or task in hand.
3. *AMOWA-UI*: we discuss an implementation of a Web user interface, which is a client-site implementation of the AMOWA-WS.

The previous chapters describe different types of developments of the AMOWA framework to examine and demonstrate the benefits of a proposed adaptive moderation approach, while Chapter 6, through quantitative and qualitative studies evaluates the proposed framework and deals with seventh contribution of the thesis. We set up three studies using our Web service and related user interface (AMOWA-WS and AMOWA-UI).

1. First study utilized a within-subjects design in order to explore the effectiveness of adaptive faceted ranking and facet selection strategies. The results of this study are divided into two parts: (1) the quantitative assessment which measures the performance using Mean Average Precision (MAP) and compares with the performance of default ranking setting (reverse-chronological ranking). (2) the subjective assessment which asks evaluators to answer questions regarding effectiveness, efficiency, and satisfaction of using such a system.
2. Second study evaluates the performance of clustering comments along different semantic facets and proposed semantic enrichment methods.
3. Third study evaluates which topic-identification algorithm is most appropriate for short texts. This helps us to define an appropriate method for identification of topics which can be used as a facet.

The first study evaluates the framework through a user study, while the second and third study investigate particular aspects of the framework. Details of the evaluation of the framework and results are under review for a publication [Momeni et al., 2014a] (to be titled “Multi-faceted Adaptive Ranking of Social Media Comments”).

Finally, in Chapter 7 we conclude our work with a qualitative analysis of our approach and discuss future research directions based on the results of this thesis.

Chapter 2

State-of-the-art-analysis

2.1 Introduction

This section aims to explore and shed light on the varying notions of “value” across different types of UGC by presenting a unifying scheme that includes the commonly used definitions of value in existing research. This is achieved by answering the following general research questions: What are the values expected to be maximized for different application domains? How are they defined with regard to the particular application domain and the task in hand? What methods are used for assessing and ranking UGC? Which methods are effective for maximizing the value of UGC with regard to an application domain? What are the effective features and metrics used to predict and measure the particular value of UGC?

The findings of a systematic review of existing approaches and methodologies for assessing and ranking UGC are presented to answer these questions. The focus is, in particular, on the short, text-based user-generated content typically found on the Web. The discussion and results of this chapter are under review for a journal article [Momeni and Cardie, 2014] and were partially published as an article [Momeni, 2012].

It is observed that the existing approaches generally adopt one of three frameworks:

1. ***Community-Based Assessment and Ranking of UGC:*** Approaches that fall under this framework use a variety of methods to classify, cluster, and rank UGC based on the majority preferences of the community. These approaches aim to maximize performance with respect to a single, pre-determined definition of value. Examples include distinguishing helpful vs. non-helpful product reviews, clustering relevant tweets according to the topic, classifying useful and non-useful comments on social media objects (e.g. YouTube videos, Flickr photos), or identifying credible postings in online forums — all based on the majority vote (or agreement) across the applications or authors of the platforms. It is observed that the proposed methods for community-based assessment and ranking approaches belong to one of two general types:
 - ***Crowd-based Methods:*** the most common method for ranking and assessing user-generated comments simply allows all users to vote on (and possibly assess and rank) the contributions of others, for example voting “thumbs up” and “thumbs down” on the comments on a YouTube video or helpfulness voting on product reviews in Amazon.
 - ***Machine-based Methods:*** the alternative method for assessing and ranking the value of user-generated comments employs a machine-learning approach such as supervised learning, unsupervised, etc. Very generally, machine-based methods first specify what is considered as valuable UGC for the application domain of interest. This may be done explicitly by providing examples of content which are valuable or not valuable. Then, a classification model is trained or a clustering method is developed to assess and rank content with regard to the defined “value”. A classification model, for example, can be trained to identify non-helpful product reviews.
2. ***Single-User Assessment and Ranking of UGC:*** These approaches aim to accommodate individual differences in the assessment and ranking of UGC through adaptive and interactive methods that personalize the results, affording an individual user the opportunity to explore content, specify the user’s own notion of “value” or interact with the system to modify the dis-

play of rankings and assessments in accordance with preferences expressed, behaviors exhibited implicitly, and details explicitly indicated by individual users. Examples include using the geo-location of a user's Twitter posts to provide neighborhood-specific information and using the content of recent posts to alert users to additional topic-relevant content on their Twitter feed [Hu et al., 2013]. These approaches can be categorized in two main groups:

- ***Personalized approaches*** assess and rank UGC according to the user's previous activities, provided content, and behaviors exhibited implicitly or explicitly to assess and rank the content.
- ***Interactive & Adaptive approaches*** do not explicitly or implicitly use a user's previous activities and provided content to assess and rank the content, but they give users opportunities to interact with the system and explore the ranked content in order to find content with regard to their particular requirements.

3. ***Incentivizing High-quality Content:*** Methods from this somewhat orthogonal framework aim to allocate available viewer attention among the user-generated contributions by finding a mechanism that both incentivizes high-quality contributions and maintains a high level of participation.

Survey Methodology and Scope: In the realization of this study, a survey is performed. First, by using popular digital library search services (such as ACM Digital Library¹ or IEEE Xplore Digital Library²), we collected articles related to assessment and ranking methods of UGC based on their titles and main keywords. These articles were published in the most influential and pioneer proceedings and Journals (such as proceedings of the international ACM WWW conference on World Wide Web, proceedings of the international ACM SIGIR conference on Research and Development in Information Retrieval, proceedings of the international AAAI conference on Weblog and Social Media, proceedings of the international conference on Web Search and Data Mining, proceedings of the SIGCHI conference on Human

¹<http://dl.acm.org>

²<http://ieeexplore.ieee.org/>

Factors in Computing Systems, etc). Second, for each relevant article which has been retrieved, all relevant articles which had been cited were collected. Third, the most relevant articles are filtered based on reviewing their abstracts and discussion sections, resulting in the retrieval of a corpus of 65 relevant articles published between 2003 and 2013 (Appendix1 lists collected articles). The approaches proposed by these articles are compared in detail and unified with respect to commonly defined values expected to be maximized and utilized methods. Fourth, the systematic review procedures described by [Kitchenham, 2004] are adhered to in conducting the survey. With regard to systematic review procedures, an attempt is made to first identify the contribution of a joint conceptualization which comprises the various approaches already developed in the field and unaddressed problems. Then, a synthesis of a new idea is created to address these problems. However, the advantages and disadvantages of these various approaches are not compared. Furthermore, a comprehensive list of the features which are effective for machine-based methods is presented.

The scope of this survey encompasses comparing and analyzing UGC assessment approaches related to the following application domains: product reviews, questions and answers in Q&A, postings and discussions in micro-blogs and forums (e.g., Twitter, online forum), and tags on social media objects (e.g. photos in Flickr or Youtube video). It is worth noting that all research on assessment of users who provide content, roles in online communities, and non-textual user-generated content (such as photos, video, etc) is excluded from the review process. The main focus of this study is the assessment and ranking of textual user-generated content on the Web.

2.2 Notion of value, expected to be maximized by assessment and ranking approaches

The assessment of UGC primarily started in 2003 with regard to two application domains, namely, postings in forums and product reviews. A highly general definition of value was previously used, extracting high quality UGC from different platforms

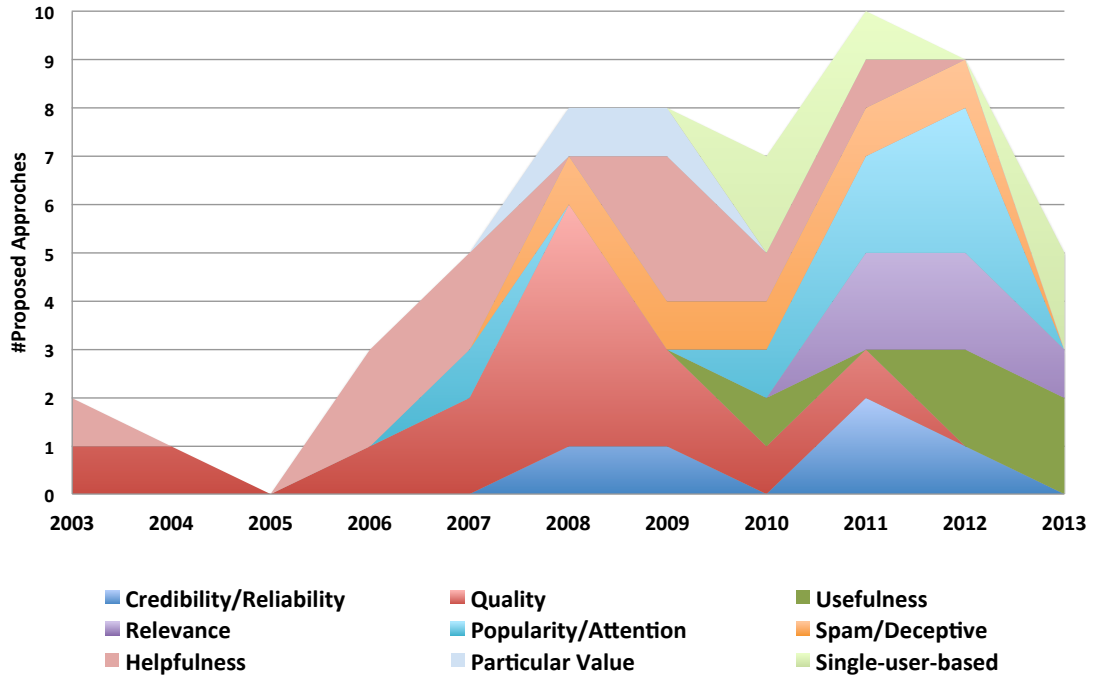


Figure 2.1: Evolution of approaches related to the assessment and ranking of UGC

and helpful product reviews. However, over time the value which has come to be expected to be maximized is defined more particularly and more sophisticatedly with more application domains being taken into consideration. Initially, quality was considered an important value. However, quality is a very general term and it has a vague definition in the context of many application domains. Therefore, the requirements to assess UGC have evolved and more dimensions of quality have become important, such as credibility, usefulness, etc. Figure 2.1 shows that evolution of approaches related to the assessment and ranking of UGC from assessment of UGC based on quality as a value, to assessment of UGC based on more sophisticated dimensions of quality (such as credibility, relevancy, usefulness, etc) as values. Furthermore, in recent years many approaches investigate in the development of single-user assessment and ranking frameworks.

In this section, the value terminologies which are accorded prime importance are described and formally defined. We also describe each value with regard to its specific definition related to a specific application domain.

Credibility is generally defined as the “quality of being convincing or believable”³. For postings in micro-blogging platforms, Castillo et al. define credibility as “credibility in the sense of believability: offering reasonable grounds for being believed” [Castillo et al., 2011]. For postings in discussion forums, Canini et al. define credibility as “credibility is associated with people who not only frequently publish topically relevant content but also are trusted by their peers.” [Canini et al., 2011]. Therefore, in order to assess the credibility of content, it is first necessary to know if the content is relevant to specific topics.

Relevance is generally defined as “closely connected or appropriate to the matter in hand”³. For postings in micro-blogging platforms, Becker et al define relevance as “relevant social media documents for a specific event” [Becker et al., 2012, Becker et al., 2011b]. Instead, for postings in micro-blogging platforms, Tao et al. define relevance as “interesting and relevant micro posts for a given topic” [Tao et al., 2012].

Usefulness is generally defined as “the quality or fact of being able to be used for a practical purpose or in several ways”³. For posting on multimedia objects (such as comments on Youtube videos), Siersdorfer et al. define usefulness as “community acceptance of new comments (community feedback for comments)” [Siersdorfer et al., 2010]. On the other hand, for an explicit definition of usefulness, Momeni et al. define usefulness as “a comment is useful if it provides descriptive information about the object beyond the usually very short title accompanying it.” [Momeni et al., 2013a]. Furthermore, Liu et al define an answer as useful in Q&A platforms “when the asker personally has closed the question, selected the best answer, and provided a rating of at least 3 stars for the best answer quality.” [Liu et al., 2007]. In the context of the micro-blogging platforms, Becker et al [Becker et al., 2011b, Becker et al., 2012] define usefulness as “the potential value of a Twitter message for someone who is interested in learning details about an event. Useful messages should provide some insight into the event, beyond simply stating that the event occurred.”. Usefulness is very closely related to helpfulness.

Helpfulness is generally defined as “giving or being ready to give help”³. Helpful-

³New Oxford American Dictionary 2011

ness is mainly defined in the product review domain and is mainly defined as the number of helpfulness votes a review received on platforms (such as Amazon.com)” [Kim et al., 2006a, Lu et al., 2010, Ghose and Ipeirotis, 2007, Jeon et al., 2006]. Juxtaposed to helpfulness in product review application domains, there are two values, namely, Spam and Deceptive. These are expected to be minimized.

Spam and Deceptive are generally defined as “giving an appearance or impression different to the true one”³. They can also be irrelevant or inappropriate messages sent on the internet to a large number of recipients. Yoo defines a deceptive product review as “a message knowingly transmitted by a sender to foster a false belief or conclusion by the receiver” [Yoo and Gretzel, 2009] and following this definition Ott et al. [Ott et al., 2012, Ott et al., 2011] define deceptive product reviews as “fictitious reviews that have been deliberately written to sound truth, to deceive the reader.”. Nithin et al. consign reviews to the category of spam when they are based upon dubious opinions and are, as a result, very damaging.

Finally, **Popularity and Attention** is “the state or condition of being liked, admired, or supported by many people”³. For postings in forums, Wagner et al. define attention as “the number of replies that a given post on a community message board yields as a measure of its attention” [Wagner et al., 2012b], [Wagner et al., 2012a], whereas Szabo and Huberman define it as “the number of votes (diggs) a story collected on Digg.com⁴ and the number of views a video received on YouTube.com” [Szabo and Huberman, 2010]. For posting in micro-blogging platforms, Hong measures popularity as the number of retweets [Hong et al., 2011].

Different application domains of UGC have different characteristics, therefore, what is defined as value varies with regard to different application domains and specific tasks in hand. Figure 2.2 shows which values are important and mainly assessed for which application domains.

⁴Digg.com is a news aggregator with an editorially driven front page, aiming to select stories specifically for the Web audience such as science, and trending political issues

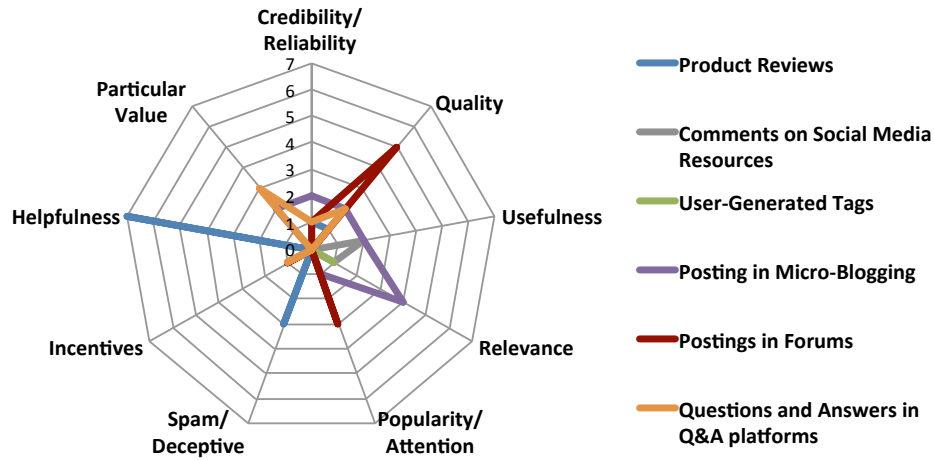


Figure 2.2: Values which are important and assessed by different application domains. Numbers on the graph indicate number of works (articles) related to each value and domain.

2.3 Influential Features Taxonomy

All social media platforms consist of three entities “*Author*”, “*User-Generated Content*”, and “*Resource*” (the media object or topic that authors generate content on). Relationships exist between these entities. Thus, for different application domains, many approaches, and particularly approaches which employ machine-based methods, utilize similar sets of features related to these entities in order to assess UGC. However, the influence of the features changes with regard to the application domain and definition of value to be maximized. In the following, a short overview of each set of features is provided. Figure 2.3 shows a taxonomy of influential features.

- ***Text-based features:*** They include characteristics founded upon aggregate statistics derived from the text of a posting, such as the readability, informativeness, average sentence length, number of punctuation marks, number of different links, and part-of-speech (POS) tagging of the words in the text.

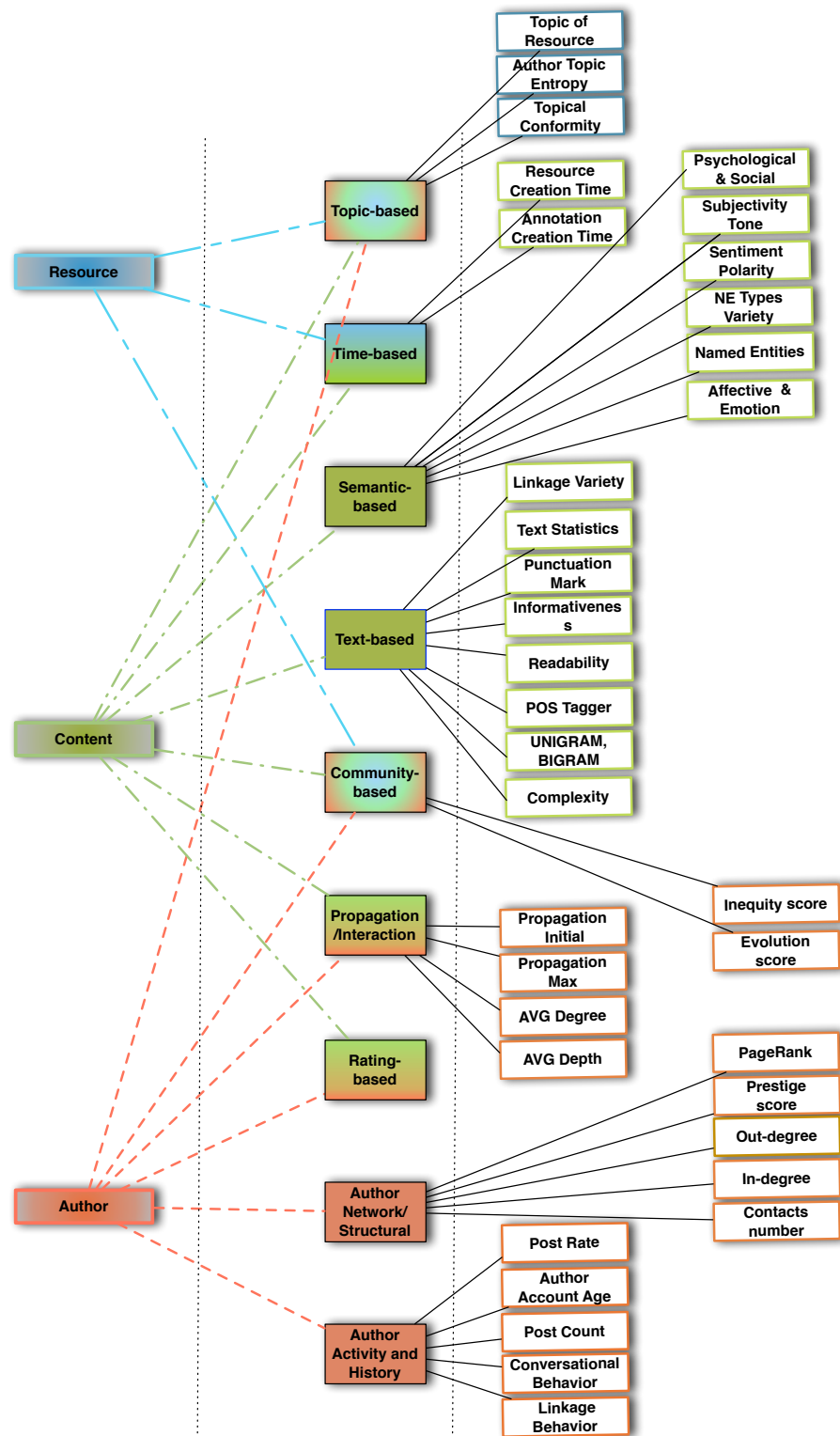


Figure 2.3: Features Taxonomy

- ***Semantic features:*** They include features related to meaning and semantics of the text of a posting, such as number of Named Entities, number of different types of Name Entities, subjectivity tone, sentiment polarity, and psychological characteristics of the content of postings.
- ***Topic-based features:*** They include standard topic modeling features that measure the topical concentration of the author of posts, topical distance of a post compared to other postings on an object, or topical distance of a post compared to other postings on a particular topic.
- ***Author activity and background features:*** These features describe the author's previous activities, behavior, and characteristics such as registration age, number of contacts (e.g., number of followers), the number of postings the author has posts in the past, and the reputation of the author (average rating that author received from the community).
- ***Author's network/structural features:*** These features capture the engagement of the author and the author's status in the social network (such as In/Out, PageRank Degree).
- ***Time-related features:*** These features are related to time, such as the time period associated with the object or topic under discussion or the time a posting was posted. For example, earlier postings may attract more attentions by community members than later postings [Szabo and Huberman, 2010].
- ***Rating-based features:*** These features are related to the rating on a post is given by a community such as average number of thumbs-up/thumbs-down or number of helpfulness votes on a posting.
- ***Community-based features:*** These include features related to relationship between content (or author) and the community with which the content is shared. For example, a user might be more likely to pay attention and reply to a post which is posted by a member of community in which the user has membership and it therefore matches topics she is interested in.

- ***Propagation/Interaction features:*** These include features related to the depth of the sharing tree and propagation tree of a posting (e.g., re-tweets).

2.4 Community-Based Assessment and Ranking of UGC

Approaches related to community-based assessment and ranking of UGC use different methods to classify, cluster, and rank UGC in accordance with the particular (a single, pre-determined) definition of the value expected to be maximized relying on majority-agreement sources of ground truth. In this section, an overview is given of these approaches with regard to three aspects: the “*value expected to be maximized*”, the “*applied method*”, and the “*application domain*”. These

The main methods proposed by the available approaches can be grouped into two categories: “*Machine-based*” and “*Crowd-based*” approaches. The majority of the available assessment and ranking approaches appear to have utilized machine-based methods for assessment and ranking of UGC. Nevertheless, the most prevalent default method utilized by many platforms is the crowd-based approach. An overview is found below which outlines available approaches related to different machine-based and crowd-based methods for different application domains and values expected to be maximized.

2.4.1 Crowd-based method

Many platforms use a crowd-based method which attempts to classify user-generated content by allowing all users to vote on the contributions of others. This wisdom-of-the-crowd approach simply allows all users to vote on (thumbs up or down, stars, etc.) or rate UGC. This method, which is also called distributed moderation, attempts to rank content according to the value estimates provided by the viewers’ votes, such as thumbs-up/thumbs-down style. Accordingly, the platforms display contributions which have attracted more votes by placing them near the top of the page and pushing those which have attracted less votes to the bottom of the page.

It is shown by Ghosh and Hummel 2011 [Ghosh and Hummel, 2011] that the crowd-based mechanism elicits extremely high quality contributions, while still achieving high participation. As a result, the lowest quality that can arise in any mixed strategy equilibrium of the crowd-based mechanism becomes optimal as the amount of available attention diverges.

Popular examples of the distributed moderation and usage of the crowd-based method are used by Yelp, Slashdot, YouTube, Reddit ⁵, and Digg. The Yelp platform permits all viewers to judge if a review written on an item is “Useful”, “Funny”, or “Cool”. The Slashdot platform is another example which filters out abusive comments by using a crowd-based moderation system. First, every comment is awarded a score of -1 to +2. Registered users receive a default score of +1, anonymous users (Anonymous Coward) receive 0, users with high “karma” receive +2, and users with low “karma” receive -1. While reading comments on articles, moderators click to moderate the comment. In addition, adding a particular descriptor to the comments such as “normal”, “off-topic”, “troll”, “redundant”, “interesting”, “informative”, “funny”, “flamebait”, etc, with each corresponding to a -1 or +1 rating, is an option for moderators. This means that a comment may have a rating of “+1 insightful” or “-1 troll”. A user’s karma increases with moderation points and a user must have a high karma to become a moderator. Being a regular user does not mean that one becomes a moderator, but instead the system gives five moderation points at a time to users based on the number of comments they have posted. In order to moderate the moderators and help reduce the number of abuses in the moderation system, the “*meta-moderation system*” is implemented. The meta-moderator examines the original comment and the arguments given by the moderator (e.g. troll, funny) for each moderation, and can judge their moderations based on the context of comments. The Youtube, the Digg, and the Reddit platforms give viewers the opportunity to judge thumbs-up/thumbs-down of comments or textual postings written on a video or article. The vote is utilized for ordering the post and discovering its place in the front-end representation. For product reviews, Amazon.com gives users possibilities to vote on the helpfulness of product reviews. More highly voted reviews are

⁵The Reddit is a news and entertainment platform where users, who register comment on submitted content (such as article and links)

displayed more prominently by placing them near the top of the page.

Lampe and Resnick [Lampe and Resnick, 2004] indicate in a summary statistic the extent to which users participate in moderation and meta-moderation systems (especially on Slashdot.com). The distribution of scores for comments, shows that the dispersal of scores for comments is reasonable and agreement on the part of the community exists on the fairness of moderations. Analyzing Slashdot.org from a statistical perspective confirms the validity of the concept which underlies distributed moderation. Users participate widely and frequently and an almost unanimous consensus is found with regard to whether a comment is moderated up or down. The dispersal of comment scores enables viewers to access potentially valuable information. On a closer analysis, it is however revealed that identifying comments may require considerable time, especially for valuable comments. Also, comments which have been incorrectly moderated are often not reversed, and comments which have low starting scores are often not treated by moderators in the same manner as other comments are. Thus, it is important to take into consideration how timely the moderation is, how accurate or inaccurate the moderation is, how influential individual moderators are, and how the input on the part of individual moderators can be reduced.

2.4.2 Machine-based method

This method employs a machine-learning approach —such as classification, clustering, etc — by precisely defining what is considered as valuable UGC for the application domain of interest. Examining machine-based methods more closely shows that many available machine-based assessment approaches use and include crowd judgments on the content in order to create a ground truth (For example, many assessment approaches for classification of product reviews with regard to helpfulness as the value have use crowd votes — helpfulness votes — to create the helpfulness ground-truth) while others due to various biases arising from judgments completely exclude crowd. It is important to note that by describing the approaches which include crowd judgments, we mean the judgments received from the internal community that the content has been shared with. Nevertheless, many approaches

use external crowd — using crowd-sourcing platforms — which are independently judged content without direct access to the source of content. Therefore, we define these approaches as approaches which exclude crowd, and this implies that they exclude internal crowd judgments.

With regard to different application domains and values, various machine learning approaches (supervised, semi-supervised, and unsupervised learning) are appropriate. For example, many assessment approaches for identifying relevant micro-blogging posts to a topic use unsupervised learning approaches, while many approaches related to assessment of UGC credibility use supervised learning methods. Furthermore, it is observed that machine-based methods for different application domains use similar sets of features (see Section 3) related to different entities of social media platforms (Author, Content, and Resources) in the particular domain in order to assess UGC. However, the influence of the features changes with regard to the application domain and definition of value to be maximized. An overview follows of approaches which use a machine-based method for assessment and ranking of UGC. For each value, first, we describe the main observations from all available works, and, second, we give a short overview of detail methods used by each work for different application domains.

Approaches for Assessing Credibility or Reliability

Examining approaches for assessing credibility or reliability more closely indicates that most of the available approaches utilize supervised learning and are mainly based on external sources of ground-truth [Castillo et al., 2011, Canini et al., 2011]. Features such as author activities and history (such as bio of a author), author network & structure, the propagation (such as a re-sharing tree of a post), and topical-based affect source credibility [Castillo et al., 2011, Morris et al., 2012]. Castillo et al. [Castillo et al., 2011] and Morris et al. [Morris et al., 2012] show that text and content-based features are themselves not enough for this task. Also, Castillo et al. indicate that authors' features are by themselves inadequate. Moreover, conducting a study on explicit and implicit credibility judgements, Canini et al. [Canini et al., 2011] find that the expertise factor has a strong impact on judging

the credibility, while social status has less impact. Based on these findings, it is suggested that in order to better convey credibility, improving the way in which social search results are displayed is required [Canini et al., 2011]. Besides, Morris et al. suggest that information regarding credentials related to the author should be readily accessible (“accessible at a glance” [Morris et al., 2012]) due to the fact it is time-consuming for a user to search for them. Such information includes factors related to consistency (such as number of posts on a topic), ratings by other users (or re-sharing or number of mentions), and information related to author personal characteristics (bio, location, number of connections).

For postings in micro-blogging platforms Castillo et al. study the information credibility of news propagated through Twitter. Detection of credible tweets is achieved by using supervised learning methods trained by manually labeled training examples. First, each tweet is labeled by a group of human annotators according to whether it corresponds to some newsworthy information or an informal conversation. Second, each tweet is assessed with regard to its level of credibility by another group of judges. Canini et al. investigate various factors which influence both explicit and implicit credibility judgments about a author and sources of reliable information in micro-blogging platforms. They propose a ranking algorithm which takes into consideration a basic text-based and authors’ social structure & network features and then ranks a list of credible authors and sources of information for a given topic. In addition, Morris et al. by conducting two controlled experiments explore users’ perceptions of tweet credibility. Several features of tweets are systematically altered to assess their impact on credibility judgments.

For questions and answers as an application domain, Bian et al. [Bian et al., 2009] propose a semi-supervised approach for assessing content credibility and author reliability which is based on coupled mutual reinforcement framework that requires only a very small number of trained samples. The proposed framework elaborates on the mutual reinforcement between the connected entities (beginning with a set of known labels for two entities, authors or answers) in each bipartite graph to assess the credibility and reputation. They state the mutual reinforcement principle as follows: *“An answer is likely to be of high quality if the content is responsive and well-formed, the question has high quality, and the answerer is of high answer-*

reputation. At the same time, a author will have high answer-reputation if she posts high- quality answers, and high question-reputation if she tends to post high-quality questions. Finally, a question is likely to be of high quality if it is well stated, is posted by a author with high question reputation, and attracts high-quality answers.” [Bian et al., 2009].

Approaches for Assessing Usefulness

Many approaches related to usefulness, use a supervised learning method to classify useful from non-useful content on social media objects [Momeni et al., 2013a, Siersdorfer et al., 2010]. These approaches show that what counts as useful content can depend on a number of factors including the practical purpose in hand, the media type of the resource (e.g., if the object is a document, video, art object, photo etc.), topic type of the resource (e.g., if the video which is commented on is associated with a person, place, event etc.), the time period associated with the resource (e.g., it is about 20th century or the 1960’s), or even the degree of opinion polarity around the resource.

For comments on social media resources (YouTube videos, Flickr photos, etc) as an application domain, semantic and topic-based features play an important role in the accurate classification of usefulness comments and especially important are those features that capture subjective tone, sentiment polarity, and the existence of named entities [Momeni et al., 2013a]. In particular, comments that mention named entities are more likely considered to useful, while those that express the emotional and affective processes of the author are more likely considered to be non-useful. Similarly, terms indicating “Insight” (e.g., think, know, consider, etc.) are associated with usefulness, while those indicating “Certainty” (e.g., always, never, etc) are associated with non-useful comments. With regard to different topics of media objects — people, places, and events — the classifier more easily recognizes useful comments for people and events regardless of the social media platform [Momeni et al., 2013a]. Therefore, training “topic-type-specific” usefulness classifiers generally allow improved performance over the “type-neutral” classifiers [Momeni et al., 2013a]. In addition, negatively rated comments by crowd which are considered as non-useful



Figure 2.4: Videos with low (lower row) versus high (upper row) variance of comment ratings. The figures are taken from [Siersdorfer et al., 2010].

content [Siersdorfer et al., 2010] contain a significantly larger number of negative sentiment terms. Similarly, positively rated comments which are considered as useful content contain a significantly larger number of positive sentiment terms [Siersdorfer et al., 2010]. Therefore, all these results suggest that for prediction of useful posts, having access to the post text-based and semantic-based features is important for this task [Paek et al., 2010, Momeni et al., 2013a].

Siersdorfer et al. [Siersdorfer et al., 2010] and Momeni et al. [Momeni et al., 2013a] propose a classifier for the curation of useful comments on social media resources such as YouTube videos. Momeni et al. construct a crowd-sourced gold standard of useful and non-useful comments and use a logistic regression model to develop a “usefulness” classifier, while Siersdorfer et al. analyze correlations between views, comments, comment ratings, and topic categories. Based on this analysis, a classifier is trained — using the support vector machine classification — to predict community usefulness assessment of comments by using community feedback on already rated comments. In addition, they make use of the publicly available SentiWordNet [Baccianella and Sebastiani, 2010] thesaurus to study the connection between sentiment scores obtained from SentiWordNet and the comment rating behavior of the community. They also study the relationship between polarizing content and comment ratings. By polarizing content they mean “content likely to trigger diverse

opinions and sentiment” [Siersdorfer et al., 2010]. In order to identify polarizing videos, the variance of comment ratings for each video is computed. (Figure 2.4 shows examples of their selected videos with high versus low rating variance. Videos about an Obama, Iraki girl stoned to death and protest on Tiananmen Square in contrast to videos about amateur music, cartoons, and The Beatles) and they show association between polarizing content and diverging and intensive comment rating behavior in Youtube [Siersdorfer et al., 2010] .

Approaches for Assessing Popularity and Attention

Many approaches related to popularity and attention, use a supervised learning method to classify content into popular (or seed) and non-popular. The temporal and author-related features are shown as important features for assessment and ranking of popular content [Hong et al., 2011, Rowe et al., 2011]. Unlike popular posts which receive tens of thousands of attentions (such as retweets, re-share), normal posts only attract a small audience and users lose interest in them very quickly [Hong et al., 2011]. Therefore, temporal features have a stronger effect on posts with a low and medium volume of attentions compared to highly popular messages. It is empirically demonstrated that the use of author features for identifying seed posts has more effect [Rowe et al., 2011] than the use of text-based features. Furthermore, it is shown that the manner in which attention is created varies with regard to different community forums. How particular features are associated positively on the start of discussions in one community may differ in another community [Wagner et al., 2012b]. Furthermore, the influential factors for predicting whether a discussion begins around a post may vary depending on the factors that impact the how long the discussion lasts [Wagner et al., 2012b, Wagner et al., 2012a]. Therefore, in forums, Wagner et al. show that the ignorance of a user is not advantageous since understanding the behavioral patterns peculiar to individual communities is influenced by posts which attract a community and stimulate long dialogues in a forum [Wagner et al., 2012a]. Also, features related to authors’ activities and history play an important role in how popular a post is [Hsu et al., 2009]. The social network provided by the service does not influence users to look at the content once content has become visible to a huge number of viewers [Szabo and Huberman, 2010], while

Table 2.1: Sample of tweets with high retweets. Samples were taken from [Hong et al., 2011].

<i>“RT @paramore Watch the World Premiere of Paramore’s new video for Brick By Boring Brick’ #paramore”</i>
<i>“RT@CamaroWRX: http://bit.ly/794Edz because everyone #needsmore-bradley”</i>
<i>“RT@narendra: Please RT. Some recent thoughts on the empathic web. that made the Huffington Post - http://bit.ly/9WyrnT”</i>

during situations with a low number of viewers they are still important. Therefore, when no early click-through information is available, predictions based on a semantic analysis of content is more useful [Szabo and Huberman, 2010].

For postings in micro-blogging platforms as an application domain, Hong et al. [Hong et al., 2011] investigate how the popularity of posts may be forecast by measuring the number of future re-sharing (such as retweet, Table 2.1 shows sample of tweets with high retweets [Hong et al., 2011]). They discuss what kinds of factors influence information propagation in Twitter. The problem of forecasting the popularity of posts is divided into two categories, both of which present classification problems: (1) a binary classification problem that predicts whether or not a message will be retweeted, and, (2) a multi-class classification problem that predicts the volume of retweets a particular message will receive in the near future. Rowe et al. [Rowe et al., 2011] present an approach to identify the characteristics of UGC that generates a high volume of attention. For predicting the superficiality or depth of a discussion, they examine the influence of text-based and author-based features for predicting the level of discussion in micro-blogging platforms based on a proposed behavior ontology (which is used to model statistical features that are used by prediction models).

For postings in online forums, Wagner et al. gain insight into behavior which is peculiar to individual community forums and which they use publicly to attract notice [Wagner et al., 2012b, Wagner et al., 2012a]. They propose an approach which is a two-step process that operates by (1) identifying posts that may obtain a reply (seed posts), and (2) predicting how much attention seed posts may create. Hsu et al. [Hsu et al., 2009] also propose a learning function based on a regression model

by collecting feedback from the community for classifying popular comments in an automatic manner. They study comments on news postings on Digg. For training the ranking function, several factors are analyzed including the author's reputation, the intricateness and informativeness of the comment, and comment visibility. Comment visibility is measured by two factors: (1) the rating of the article by the community, and (2) the time of posting the comment. With regard to the rich-get-richer visibility cycle, they propose re-scaling the ratings of comments for each training sample. So, a comment with a huge number of ratings, which is placed in a low average rating area and has small variance, is advanced to a new boosted rating.

Finally, Szabo and Huberman [Szabo and Huberman, 2010] propose an approach for predicting the long-term popularity of UGC based on early assessment of user access. Based on experiments on two well-known social media platforms, Digg and YouTube, they show that in Digg, assessment of access to given stories during the first two hours after posting enables us to estimate their popularity in within the next 30 days with a relative error of 10%, while predicting the polarity of YouTube video (with regard to download rate of YouTube videos) needs to be followed for 10 days to achieve the same relative error. The influence of time on predictions is due to differences in how content is consumed on the two platforms. Posts on Digg become outdated very fast, while posts on YouTube videos become outdated much later. Therefore, predictions are more accurate for content with a short life cycle, whereas for predictions for content with a longer life cycle, greater statistical error is more likely.

Approaches for Assessing Helpfulness

Helpfulness is mainly discussed in the product reviews domain. This is due to the fact that many online shopping and online booking platforms explicitly ask their users to vote on the helpfulness of product reviews (Figure 2.5 shows example of helpfulness voting system of Amazon online shopping service⁶). Accordingly, many machine-based approaches utilize and learn from these votes to train and develop a model. Many of the available approaches use supervised learning methods based

⁶<http://www.amazon.com/>

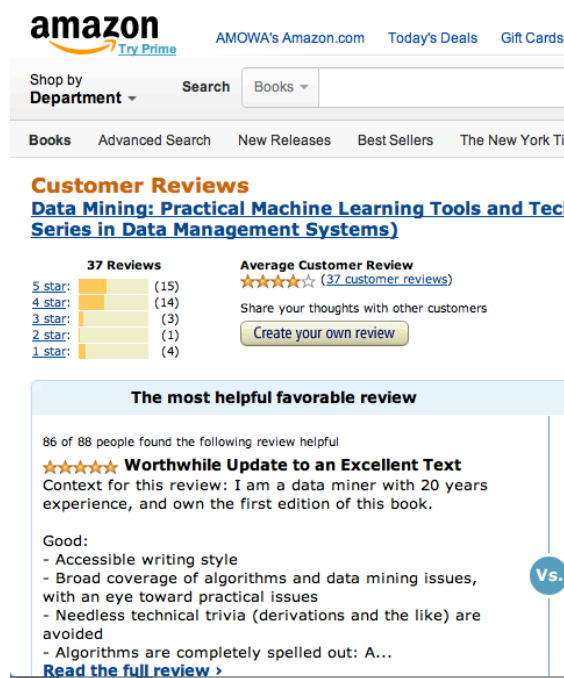


Figure 2.5: Helpfulness voting system of Amazon online shopping service (amazon.com).

on user votes as the ground-truth [Kim et al., 2006a, O’Mahony and Smyth, 2009]. However, there are few works based on the semi-supervised [Lu et al., 2010] and supervised learning [Tsur and Rappoport, 2009].

Many 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 include: length of the review [Kim et al., 2006a], mixture of subjective and objective information, readability such as checking the number of spelling errors, conformity (the helpfulness of a review is greater when the star rating it has received is more similar to the aggregate star rating of the product) [Danescu-Niculescu-Mizil et al., 2009, Kim et al., 2006a], and author reputation and social context features [O’Mahony and Smyth, 2009, Lu et al., 2010]. However, the effectiveness of features related to social context depends on there being sufficient training data to train these extra features [Lu et al., 2010], and features related to social context are less successful in comparison to author reputation features [O’Mahony and Smyth, 2009]. Furthermore, it is demonstrated that helpfulness of

a product review is based on properties actually found in the review itself and is not necessarily consistent with its similarity to the corresponding product description [Zhang and Varadarajan, 2006] and it is shown that the helpfulness a product review is considered to have a slight correlation with the subjectivity or sentiment polarity of a review’s text [Zhang and Varadarajan, 2006].

With regard to supervised learning, Mahony and Smyth propose a recommender system based on the classification method that is designed to recommend the helpful reviews for a given product [O’Mahony and Smyth, 2009]. First, they train a classification approach using users’ feedback. The proposed approach is evaluated using a large set of TripAdvisor⁷ hotel reviews. Prediction confidence scores are then used to effectively rank reviews. This is carried out by ordering those reviews classified as helpful with regard to the prediction confidence. Zhang and Varadarajan [Zhang and Varadarajan, 2006] based on a regression model propose a method for predicting utility, reliability, helpfulness, and informativeness of product reviews. To rank helpful product reviews two methods are proposed by Ghose et al. [Ghose and Ipeirotis, 2007, Ghose and Ipeirotis, 2011]. A “consumer-oriented ranking mechanism” orders the reviews in accordance with the helpfulness which had been anticipated, while a “manufacturer-oriented ranking mechanism” orders the reviews in accordance with the effect on sales which had been anticipated. This method integrates subjectivity and econometric analyses. Also, Kim et al. [Kim et al., 2006a] study how predicting product review helpfulness can be carried out automatically by exploiting helpfulness votes on Amazon.com and use helpfulness votes as ground-truth to train a helpfulness function.

So far, many proposed approaches have utilized users-ratings for developing and training a prediction model. However, Liu et al. [Liu et al., 2007] show that users-ratings at Amazon have three kinds of biases. These are: (1) “imbalance vote bias”, (2) “winner circle bias”, and (3) “early bird bias” [Liu et al., 2007]. Therefore, they propose a specification — a guideline for what a good review consists of to measure the quality of product reviews — and propose a classification-based approach developed from manually annotated product reviews which complies with

⁷TripAdvisor.com is a travel website providing directory information and reviews of travel-related content

the proposed specification. The proposed approach explores three aspects of product reviews, namely informativeness, readability, and subjectiveness. The results show that both the features on word level and those on product feature level can improve the performance of classification significantly. The features on readability can increase the accuracy, but their influence is considerably less [Liu et al., 2007].

With regard to semi-supervised learning methods, Lu et al. [Lu et al., 2010] exploit information gleaned from texts about authors' identities and social networks for predicting helpfulness of product reviews. They propose a semi-supervised approach for exploring social context information by adding regularization constraints to the linear text-based predictor. Four constraints are defined: (1) "Author Consistency", (2) "Trust Consistency", (3) "Co-Citation Consistency", and (4) "Link Consistency" [Lu et al., 2010]. Two different methods are explored for incorporating the social context information into the helpfulness predictor model. The first method extends the feature space in a straightforward manner by adding features extracted from the social context. The second method utilized defined constraints between reviews and reviewers, and then integrates regularizers to the linear regression formulation to apply these constraints.

With regard to unsupervised learning methods, Tsur and Rappoport describe a method for ranking helpful product reviews which contrasts with many of the proposed supervised learning approaches [Tsur and Rappoport, 2009] — "REVRANK" algorithm. "REVRANK" algorithm first created a virtual optimal review by identifying a core of dominant words found in reviews. This is achieved in two stages. First, identification of dominant words by how often they are used, and then the identification of words that are used less often, but provide pertinent information on the specific product. Second, using these words, a definition of the "feature vector representation" of the most desired review is created. Finally, reviews are transposed to this representation and ordered with regard to their similarity from the "virtual core" review vector.

Finally, Danescu et al. [Danescu-Niculescu-Mizil et al., 2009] analyze the evaluation of opinions on product reviews by exploiting the phenomenon of review plagiarism. They show that helpfulness votes on online platforms provide a way to assess how helpfulness ratings are evaluated by members of an on-line community on a very

large scale. In addition, they show that the perceived helpfulness ratings correlate with other evaluations of the same product of a review and not necessarily with the content of reviews.

Approaches for Assessing Spam and Deceptive

Similar to helpfulness, assessing spam and deceptive content is mainly discussed in the product reviews domain. Approaches in these areas can be basically categorized into two groups: (1) approaches for assessing spam product reviews [Jindal and Liu, , Jindal and Liu, 2008] (Product reviews on brands, duplicates, and non-reviews such as advertisements, other irrelevant reviews), (2) approaches for assessing deceptive product reviews [Ott et al., 2012, Ott et al., 2011]. Deceptive product reviews are which are written so that the reader believes they contain the truth, but instead they actually give the reader false information [Ott et al., 2012, Ott et al., 2011]. Approaches related to both groups utilize supervised learning methods and mainly use text and content related features. For assessing spam product reviews, three types of features are used: (1) “review centric” features, which include rating-based and text-based features, (2) “reviewer centric” features, which author-based features, and (3) “product centric” features [Jindal and Liu, , Jindal and Liu, 2008]. The highest accuracy is achieved by using all the features. However, the model performs as efficiently without using rating-based features. Rating-based features are not effective factors for distinguishing spam and non-spam because rating (feedbacks) can also be spammed [Jindal and Liu, , Jindal and Liu, 2008]. For assessing deceptive product reviews, n-gram related features have the highest impact, but an approach which combines psycho-linguistically related features and n-gram features can achieve slightly improved results. Moreover, there is a reasonable correlation between deceptive opinion and imaginative writing based on similarities of distributions of Part of Speech tags [Ott et al., 2011].

With regard to approaches for assessing spam product reviews, Jindal and Liu [Jindal and Liu, , Jindal and Liu, 2008] study opinion spam in product reviews. The main goal of this work is to detect and rank spammed product reviews. Spam is defined as belonging to three categories: (1) opinions which are not based on the truth,

Table 2.2: Sample of deceptive and truthful reviews. Samples were taken from [Ott et al., 2011].

Review	Type
<i>“I have stayed at many hotels traveling for both business and pleasure and I can honestly say that The James is tops. The service at the hotel is first class. The rooms are modern and very comfortable. The location is perfect within walking distance to all of the great sights and restaurants. Highly recommend to both business travellers and couples.”</i>	Truthfull
<i>“My husband and I stayed at the James Chicago Hotel for our anniversary. This place is fantastic! We knew as soon as we arrived we made the right choice! The rooms are BEAUTIFUL and the staff very attentive and wonderful!! The area of the hotel is great, since I love to shop I couldn’t ask for more!! We will definatly be back to Chicago and we will for sure be back to the James Chicago.”</i>	Deceptive

(2) reviews which deal with proprietary names, and (3) “non-reviews” (such as advertisements or other irrelevant reviews) [Jindal and Liu, , Jindal and Liu, 2008]. Detection of type 2 and type 3 spam reviews is conducted by using supervised learning based on manually labeled training samples of reviews. Detection of type 1, untruthful opinions, is carried out by detecting duplicate reviews (*“duplicates from different user-ids on the same product, duplicates from the same user-id on different products, and duplicates from different user-ids on different products”* [Jindal and Liu, 2008]). To train a prediction model, duplicate spam reviews are utilized as positive training samples, while other reviews are utilized as negative training samples.

With regard to approaches for assessing deceptive product review, Ott et al. study opinion spam in product reviews with specific focus on types of opinion spam which could be very dangerous and misleading [Ott et al., 2012, Ott et al., 2011]. Such fabricated opinions are intentionally composed to appear genuine in order to trick the reader (Table 2.2 shows sample of deceptive and truthful product reviews). Their study follows the Yoo and Gretzel [Yoo and Gretzel, 2009] approach for comparing the syntax of misleading and truthful hotel reviews. They also present a general framework for forecasting the incidence of deception in online review communities. Recognition of deceptive spam reviews is done by using supervised learning

with manually labeled training samples. For collecting deceptive reviews they obtained coders from the Amazon’s Mechanical Turk Service ⁸. Coders were asked to write a review which would favorably advertise the hotel. They were told that the review should be persuasive and matter-of fact. Furthermore, Yoo and Gretzel [Yoo and Gretzel, 2009] present a study which was conducted to compare how deceptive and truthful hotel reviews are constructed linguistically. The results show that deceptive and truthful reviews vary with regard to the complexity of vocabulary, personal and impersonal use of language, trademarks, and personal feelings. Nevertheless, the results tend to indicate that linguistic features of a text are simply not enough to distinguish between false and truthful reviews [Yoo and Gretzel, 2009]. In order to prepare a training set, students of tourism marketing were asked to write a review which would positively advertise the hotel.

Approaches for Assessing Relevance

Most of the available approaches related to relevance are based on unsupervised learning.

For postings in micro-blogging platforms as an application domain, Becker et al. [Becker et al., 2011b, Becker et al., 2012] explore approaches for finding representative posts among a set of Twitter messages that are relevant to the same event. Their aim being to identify high quality, relevant posts that provide useful information about an event. The problem is approached in two concrete steps. First, by identifying each event and its associated tweets using a clustering technique that clusters together topically similar posts. Second, for each cluster of event, posts are selected that best represent the event. Centrality-based techniques are used to identify relevant posts with high textual quality and are useful for people looking for information about the event. Quality refers to the textual quality of the messages — how well the text can be understood by any person. From three centrality-based approaches (Centroid, LexRank [Radev, 2004], and Degree), Centroid is found as the preferred way to select tweets given a cluster of messages related to an event [Becker et al., 2012]. Furthermore, Becker et al. [Becker et al., 2011a] investigate

⁸mturk.com is an online crowd sourcing service

approaches for analyzing the stream of tweets to distinguish between posts about “real-world” events and “non-event” messages. First, they identify each event and its related tweets by using a clustering technique that clusters together topically similar tweets. Then they compute a set of features for each cluster to help determine which clusters correspond to events and use these features to train a classifier to recognizing between event and non-event clusters. In the same application domain, Tao et al. [Tao et al., 2012] explore if additional micro-post characteristics exist that are more predictive of the relevance of a post rather than of its keyword-based similarity for quering in micro-blogging platforms such as Twitter. They investigate sixteen features along two dimensions: “topic dependent” and “topic-independent” features.

For question and answer postings as an application domain, Bian et al. by investigating user interactions, rating-based and community-based features present a ranking framework to find high quality, relevant questions and answers with factually correct and well-formed content in Q&A platforms and they take advantage of information related to user interaction for building the ranking model [Bian et al., 2008]. They independently label a number of answers manually in order to evaluate the accuracy of the predicted relevance labels. Their findings show that textual, community, and user feedback (while they are noisy) features are important to improve the training of the ranking functions.

Approaches for Assessing Particular (Unique) Value

For some domains, especially in the Q&A platforms, there are values which are not examined in the majority of assessment approaches, but which are nevertheless the focus of an approach. Among these works in the Q&A domain, there are approaches for distinguishing between posts such as editorials from news stories, subjective from objective posts, or the conversational from informational posts.

For distinguishing between question and answer postings with a very large number of opinions written about current events, Yu and Hatzivassiloglou present a classifier [Yu and Hatzivassiloglou, 2003]. They show that at document level, a Bayesian classifier can differentiate between “factual” and “opinion” posts by using lexical

information. Instead, the task is significantly more difficult at sentence level. Furthermore, features such as words, bigrams, trigrams, polarity, and part-of speech play an important role for this task [Yu and Hatzivassiloglou, 2003].

For predicting a question’s subjectivity or objectivity in a Q&A site, Li et al. [Li et al., 2008] present the “CoCQA” model which is based on the concept of co-training [Blum and Mitchell, 1998] (semi-supervised learning approach). It is expected that objective questions are answered with well-founded information. Instead, subjective questions result in answers belying personal, emotional states. For creating an experimental dataset, they download questions from every top-level category of Yahoo! Answers and randomly chose a set of questions from each category to be labeled by coders from the Amazon’s Mechanical Turk Service. With regard to the feature set, they compute question and answer content and three term weighting schemes separately (such as Binary, TF, and TF-IDF⁹). By applying “CoCQA” to this task, they show they can significantly decrease the amount of the required training data.

For distinguishing between “conversational” questions and “informational” questions in Q&A platforms, Harper et al. [Harper et al., 2009] propose a classifier. They define conversational questions and informational questions as follows: *“Informational questions are asked with the intent of getting information that the asker hopes to learn or use via fact- or advice-oriented answers. Conversational questions are asked with the intent of stimulating discussion. They may be aimed at getting opinions, or they may be acts of self-expression.”* [Harper et al., 2009]. They develop an online coding tool and use data from three well-known Q&A sites (Yahoo Answers, Answerbag, and Ask Metafilter) for human coding. Based on their human coding evaluation, they show that people are able to reliably differentiate between questions which are part of a conversation and questions which ask for information and demonstrate that the archival value of the former is lower than that of the latter. For training a classifier, they evaluate several structural properties and features related to the social network model. They show that features related to structure of the text are important for distinguishing conversational and informational questions. With regard to the social network features, they show that none of these

⁹Term Frequency–Inverse Document Frequency

Table 2.3: Samples of Request and Introduction posts. Samples were taken from [Burke et al., 2007].

Post	Type
<i>“I was recently diagnosed with Epilepsy. I’ve had what I thought were Ôpanic at- tacksÕ for several years, mostly since the teen years, but it turns out they have been various types of seizures”</i>	Introductions
<i>“What can I expect from chemotherapy?”</i>	Requests

features improves performance, despite there being potentially more indicators to be extracted from the text [Harper et al., 2009]. Furthermore, they show that taking into the consideration only questions is not simply enough for classifying a Q&A thread.

For postings in online forums, Burke et al. [Burke et al., 2007] by using posts from Usenet¹⁰ conduct a series of studies related to the impact of two rhetorical strategies on community responsiveness: “Introductions” and “Requests” (show a request of the author). Table 2.3 shows samples of Request and Introduction posts. They show that “Requests” attract more community responses and community responses have a higher correlation for detection of “Requests” compared to other contextual and text-based features, such as length of posts and number of posts and contributions in a group.

Approaches for Assessing Quality

Quality is a very general term which is mainly discussed in three application domains: posting in forum platforms, assessing and ranking questions and answers in Q&A platforms, and assessing high quality user-generated tags. However, the requirements to assess UGC have evolved and more dimensions of quality have become important over time.

Approaches related to assessing the quality of questions and answers show that the combination of different types of features is likely to increase the classifier’s accuracy and adding knowledge about the author is very important when assessing the

¹⁰Usenet is a worldwide distributed Internet discussion system

quality of questions or answers [Agichtein et al., 2008, Jeon et al., 2006]. However, the reputation of the authors submitting the answers is not as important as many other features. This suggests that authority, expertise, and history of author are only important for some but not all of the predictions [Liu et al., 2008].

For Finding High-Quality questions and Answers in Q&A platforms, Agichtein et al. [Agichtein et al., 2008] and Jeon et al. [Jeon et al., 2006] present a classification framework which exploits non-textual information found in social media (e.g., community feedback) to detect high quality content. Jeon et al. [Jeon et al., 2006] collect question and answer pairs (Q&A pairs) from the Naver¹¹ Q&A service and manually collect judgments on answer quality and relevancy and, by using kernel density prediction and the maximum entropy method, exploit various types of non-textual features and create a classifier. Subsequent to research by Agichtein et al. [Agichtein et al., 2008], Liu et al. [Liu et al., 2008] first introduce the challenge presented by forecasting UGC seeker satisfaction in Q&A platforms and develop a prediction model for predicting whether the author will find answers to her question satisfactory. They define information seeker satisfaction as follows: *“An asker in a QA community is considered satisfied iff: the asker personally has closed the question, selected the best answer, and provided a rating of at least 3 stars for the best answer quality. Otherwise, we define the asker to be unsatisfied.”* [Agichtein et al., 2008].

Furthermore, For postings in Q&A domain, Harper et al. [Harper et al., 2008] explore influential factors on answer quality by conducting a comparative, controlled field study of answers posted across different types of Q&A platforms: “digital reference services”, “ask an expert services”, and “Q&A sites”. “Digital reference” services enable users to access library reference services. “Ask an expert services” is manned by “experts” in different topic areas, such as science (e.g. at “MadSci Network”¹²) or oceanography (e.g. at “Ask Jake, the SeaDog”¹³). First, they show “you get what you pay for” [Harper et al., 2008]. For example, answer quality is better in Google Answers than in the free platforms, and paying more money for an answer has a positive impact on the likelihood of receiving high quality answers. Sec-

¹¹Naver.com is a popular search portal in South Korea

¹²<http://www.madsci.org>

¹³<http://www.whaletimes.org>

ond, Q&A platforms with different types of users are more successful. For example, Yahoo! Answers which is open to the public for answering questions outperforms platforms that depend on specific users to answer questions.

For posting in micro-blogging platforms as an application domain, Diakopoulos and Naaman [Diakopoulos and Naaman, 2011] explore the correlation between comment quality and consumption and production of news information. They also describe and explore what motivates readers and writers of news comments. Their results have shown: (1) how much low quality comments influence users and journalists, (2) how perceptions of quality can be influenced by various reading motivations of individual, and (3) how flagging, moderation, and engagement can be used as policies for enhancing quality. Furthermore, they show that aspects peculiar to many online communities include unpredictable participation patterns (such as interaction between regular users and other actors in different situations.).

Finally, for posting in forum platforms as an application domain, Weimer et al. [Weimer et al., 2007] and Veloso et al. [Veloso et al., 2007] present supervised approaches to assess the quality of forum posts in the online forums which learns from human ratings. Weimer et al. use the Nabble¹⁴ platform as a data source, while Veloso et al. use a collection of comments posted to the Slashdot¹⁵ forum.

For assessing high quality tags on media resources (such as online photos), Weinberger et al. [Weinberger et al., 2008] propose a method that assesses the ambiguity level of a tag set, and to supplement this method, they propose two additional tags to resolve the ambiguity. They define a tag as ambiguous if: *“A tag set is ambiguous if it can appear in at least two different tag contexts”* [Weinberger et al., 2008]. The tag contexts are defined as *“the distribution over all tag co-occurrences”*.¹⁶ They use 50 different tags (the ambiguity evaluated by users) for evaluating and examining parameters of the algorithm. They show that the majority of the ambiguous

¹⁴Nabble.com provides an embeddable forum, embeddable photo gallery, embeddable news, embeddable blog, embeddable mailing list and archive

¹⁵Slashdot.org is a news forum

¹⁶A prime example is “Cambridge”, a city found both in Massachusetts and England. A tag such as “university” makes sense if it is used in both contexts, but the ambiguity remains unresolved. Thus, in the case of the tag “Cambridge”, the method notes that this tag contains ambiguity and recommends “MA” or “UK” [Weinberger et al., 2008]

tags is found within one of three dimensions: “temporal”, “geographic” or “semantic”. Sen et al. [Sen et al., 2007] explore implicit (behavioral) and explicit (rating) feedback to analyze and devise methods for identifying high quality tags. They investigate different lightweight interfaces for collecting feedback from members about tags to identify which interfaces result in the richest metadata for determining the quality of individual tags. Implicit system usage data and explicit feedback by members are then employed to devise a method for predicting tag quality. As a result they propose guidelines for designers of tagging systems: (1) *“Use systems that both support positive and negative ratings”*, (2) *“Use tag selection methods that normalize each user’s influence”*, (3) *“Incorporate both behavioral and rating-based”*, and (4) *“Assume that a user’s rating for a particular tag application extends to other applications of the tag”* [Sen et al., 2007].

For the same domain, Sigurbjornsson et al. [Sigurbjörnsson and van Zwol, 2008] present a characterization of tag behavior in Flickr which might be useful for the tag recommendation system and evaluation. They take a random set of Flickr photos to analyze how users tag their uploaded media objects (such as photos) and what types of tags are created. Their results show that the tag frequency distribution is associated with a perfect power law, and indicate that the middle part of this distribution contains the most interesting tags, which can be used for tag recommendation systems. Furthermore, they find that generality of the photos are included with only a few tags. Finally, Hall, et al. [Hall and Zarro, 2011] compare the metadata created by two different communities, the ipl2 digital library¹⁷, and the social tagging system, Delicious.com¹⁸. Their results show that user-contributed tags from Delicious which have the potential to be used as additional access points for ipl2 digital potentially benefit from user-library resources. The intersection area between the tags applied to ipl2 resources and indexing indicates that the two groups are similar enough to be helpful, but are nevertheless dissimilar enough for new access points and description. Furthermore, Nov et al. [Nov et al., 2008] present a

¹⁷ipl2 was born as the Internet Public Library in 1995 in a library and Information Science class taught by Joe Janes at the University of Michigan, with the central motivating question of “what does librarianship have to say to the networked environment and vice-versa?”

¹⁸Delicious (formerly del.icio.us) is a social tagging Web service for sharing, and exploring Web tags.

quantitative study and examine what motivations are associated with tagging levels. They conduct a study of tagging on Flickr. They discover that two of the three motivation categories (“Self”, “Family & Friends”, and “Public”) impact users’ tagging levels. They find that the levels of “Self” and “Public” motivations, the social presence indicators, and the number of photos have positive impact on tagging level, while the “Family & Friends” motivation is found not to be significantly in correlation with the tagging level [Nov et al., 2008].

An alternative work related to assessment of quality of UGC is proposed by Laniado and Mika [Laniado and Mika, 2010]. They analyze the extent to which a hashtag can act as an identifier for the Semantic Web. By using Vector Space Model (VSM), they propose four metrics to measure this: (1) “Frequency” refers to a hashtag being used reasonably often by a community of users [Laniado and Mika, 2010] (2) “Specificity” refers to how the usage of a word may differ, depending on whether a hashtag is used or not [Laniado and Mika, 2010] (3) “Consistency” refers to the meaning that may be attributed to a word as a result of the consistent usage of a hashtag by different users in various context [Laniado and Mika, 2010] (4) “Stability over time” refers to meaning acquired by a hashtag as a result of it being used repeatedly and relentlessly over time [Laniado and Mika, 2010].

2.4.3 Summary

Figure 2.6 provides an overview of community-based assessment and ranking approaches of UGC.

With regard to which “*methods*” are applied, it is observed that most of the available approaches related to community-based ranking and assessing of UGC utilize machine-based methods. Nevertheless, default methods, which are utilized by many platforms, are crowd-based. Examining machine-based methods more closely reveals that some machine-based assessment approaches use crowd judgments on the content in order to create a ground truth, while other machine-based assessment approaches completely exclude crowd for three reasons: (1) different biases of crowd-based approaches such as “imbalance voting”, “winner circle”, “early bird”, etc. [Liu et al., 2007], (2) a lack of an explicit definition of value which may be re-

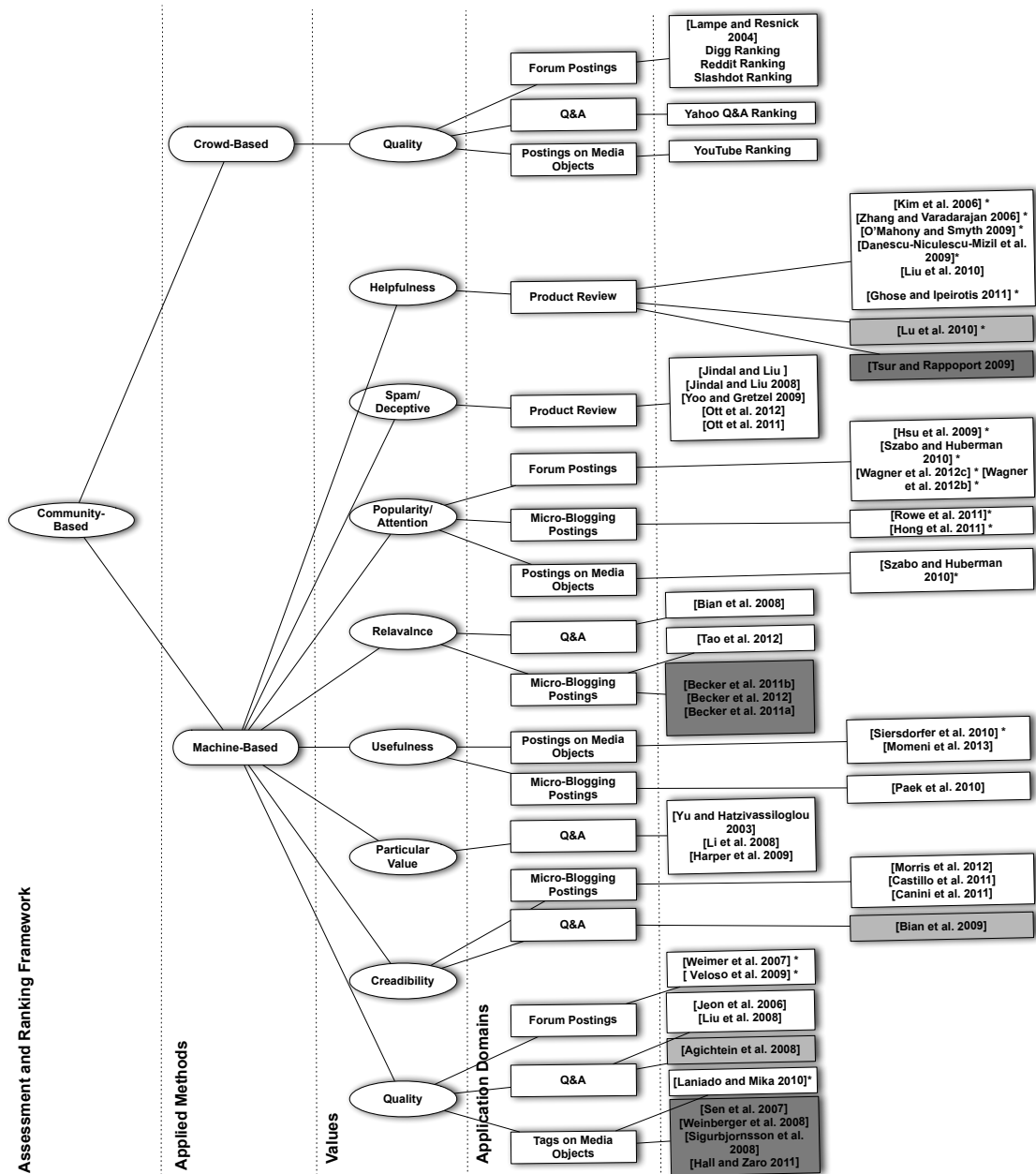


Figure 2.6: Overview of Community-Based Assessment and Ranking of UGC Approaches. At lowest level, related to citations, dark gray boxes show approaches, which utilize unsupervised learning, bright gray boxes show approaches, which utilize semi-supervised learning, and white boxes show approaches, which use supervised learning. “*” beside the citation indicate that the approach utilizes crowd as the ground-truth.

requested by the crowd to assess some application domains. For example, many assessment approaches for classification of product reviews with regard to helpfulness as the value have used crowd-based or a combination of crowd and machine-based approaches. This is because many product review platforms have explicitly defined and asked crowd to assess the helpfulness of product reviews. However, most approaches related to assessment of credibility exclude crowd-based judgments because no platforms or domains have asked the crowd for credibility judgments, and (3) human judgments can not be as precise as machine-based judgments in the case of some application domains and values, for example with regard to the exactness of truthful product reviews [Ott et al., 2012]. Approaches which exclude crowd mainly utilize two methods to create a ground-truth or training set: (1) using external crowd (using crowd-sourcing platforms) which independently judges content with regard to particular value, and (2) developing their own coding system for collecting independent judgments from a closed set of users.

With regard to the application “*domain*”, a more detailed examination leads us to discover that many proposed machine-based assessment approaches utilize supervised methods. However, approaches in the Q&A domain utilize semi-supervised learning approaches such as co-training or mutually reinforcing approaches. This is the result of the interconnectedness and interdependency between three sets of entities in Q&A (“Questions”, “Answers” and “Users”). Besides, it is observed that most of the available approaches focus on maximizing different values for micro-blogging platforms. This may be due to the very simple and structured characteristics of these platforms.

With regard to different “*values*” which are expected to be maximized, many approaches appear to maximize quality in general, applying a crowd-based method. Many approaches which aim to maximize helpfulness are mainly discussed in the domain of the product review, where crowd-judgments are predominantly used as the ground-truth to build the prediction model of these approaches. However, the use of the crowd for this value is debated. Similarly, spam and deceptive are also mainly discussed in the domain of the product review as in the case of helpfulness, but they differ from helpfulness in that they mainly exclude crowd-judgments. Also, approaches related to the assessment of popularity mainly develop their identifica-

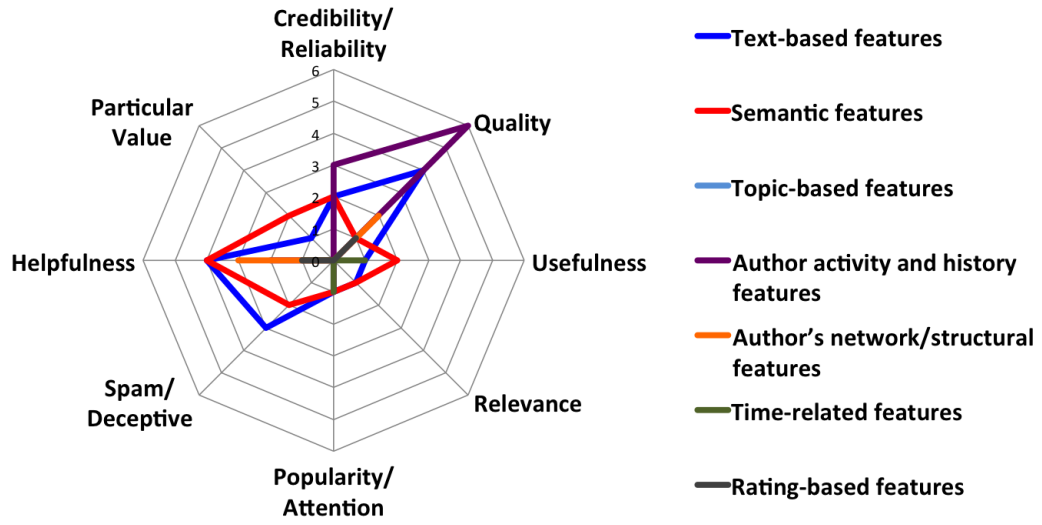


Figure 2.7: Influential features sets for assessment and ranking of different values. Numbers on the graph indicate number of works (articles) related to each value and each feature set.

tion and prediction models based on user votes and ratings of crowd (in the case of Tweeter, re-tweet). With regard to approaches related to assessment of the relevancy of UGC, many approaches employ unsupervised learning approaches due to the fact that relevancy is influenced by textual features. Therefore, applying unsupervised text clustering methods is effective for reducing the effort of labeling a huge number of unlabeled content and for maximizing this value. Finally, some features have high impact for assessment of a particular value based on our feature analysis. Therefore, for maximizing some values, systems should take into consideration the easier way to build influential features during the design phase. For example, value related to credibility should take users' profile pages into account. In the following section we give short overview of the examination of influential features for various values.

Influential Features for Different Values and Application Domains

Figure 2.7 shows influential features sets for different values with regard to proposed approaches on UGC assessment, related to classification type.

A more detailed examination of features leads us to discover that many text-based and semantic-based features are important for classifying and clustering user-generated content in all application domains. It should be noted that the features to be used depend on the notion of the value which is expected to be maximized. For quality, almost all features are helpful to achieve higher assessment accuracy because quality is in itself a very general notion. Similar to the assessment of quality, popularity requires many features to be used in its assessment and some of the more important ones include authors' activities, background, networks and structures, and propagation and interactive features. These features related to authors' activities and networks also play an important role when assessing credibility because features simply related to texts can not help to assess the credibility of postings. So, we require more contextual features to be included. However, in the case of assessing spam and deceptive content, authors can write fake reviews that have been written to appear the truth and to deceive the reader. Accordingly, the features related to the text and semantic of a review are important features to assess spam and deceptive content. Similar to the assessment of spam and deceptive content, text and semantic-based features are very often influential when assessing relevancy.

In platforms where a particular value is explicitly asked for from the crowd, rating-based features naturally play important roles for assessment of the value. An example is helpfulness of product reviews in many platforms when judgments on the helpfulness are requested from the crowd. It is worth noting that time-based features play an important role for assessing helpfulness, usefulness, and popularity. Finally, community-based features are mainly taken into consideration for assessment of postings in forums which include different communities.

2.5 Single-User Assessment and Ranking of UGC

These approaches use different methods to allow for the differences between individual users for adaptive, interactive or personalized assessment and ranking of UGC. Individual users are given the opportunity to explore content, personally define the expected value, or interact with the system to adapt the display of ranking and rank

content in accordance with individual user requirements. These approaches can be categorized in two main groups: “Personalized Approaches” and “Interactive & Adaptive Approaches”. In the following, an overview is given of these approaches.

2.5.1 Personalized Approaches

Personalization approaches assess and rank UGC relevant to the individual user, taking into account how the user acted previously, what activities she participated in, what implicit behavior and preferences can be observed, and what details were explicitly provided.

For posting in micro-blogging platforms, Burgess et al.[Burgess et al., 2013] propose “BUTTERWORTH” which is a service that helps users find more relevant content to their interest on their feeds without using explicit user input. “BUTTERWORTH” automatically generates a set of “rankers” by clustering sub-communities of the user’s contact based on the common content they produce. The proposed service comprises three main components. First, the “list generator” groups friends into lists by examining their social contact. Second, the “list labeler” generates a human-readable label representing the list’s topic. Third, the “topic ranker” trains ranking models for core topics. The models can then be utilized to order the user’s feed by the selected topic [Burgess et al., 2013]. Uysal et al.[Uysal and Croft, 2011] propose a personalized ranking of tweets by exploiting users’ retweeting patterns and conduct a pilot user study to explore the correlation between retweeting and the interestingness of the tweets for an individual user.

For posting in online forums, it is recommended by Lampe et al. [Lampe et al., 2007] that for ranking comments, patterns recognized by setting filters of users can be used to minimize the cost of settings for other users. [Lampe et al., 2007]. One suggested strategy is creating static schema that take into consideration the filtering patterns of different groups of viewers. Another strategy is the setting of filtering thresholds for each conversational tree dynamically, based on the selections of previous viewers and shows that selections previously made by readers are much more helpful than content of postings for this task (for example the ratings of those comments). Moreover, it is discovered that users can be grouped in three categories: *“those*

who never change the default comment display”, *“those who use ratings to modify the display*”, and *“those who change the comment display to suppress ratings”* [Lampe et al., 2007] and a large number of users do not change from system set default setting. Furthermore, Hong et al. [Hong et al., 2012] explore the creation of ranking systems by proposing a probabilistic latent factor model for social feeds from the perspective of LinkedIn¹⁹. Particularly, they convey this task as an intersection of learning for ranking, “collaborative filtering”, and “clickthrough modeling”.

2.5.2 Interactive & Adaptive Approaches

The term “adaptation” refers to a process in which an interactive system (adaptive system) adapts its behavior to individual users based on information acquired about its users and their environment. The main difference of these systems compared to personalized systems is that these systems do not explicitly or implicitly use users’ previous common actions and activities to assess and rank the content. However, they give users opportunities to interact with the system and explore the ranked content in order to find content with regard to their requirements.

Recently, there have been bodies of research which combine machine-based, personalized, and adaptive ranking approaches to assess UGC and maximize values for their viewers. As an example of such a system for posting in micro-blogging platforms, Hu et al.[Hu et al., 2013] propose Whoo.ly, a Web service that provides “neighborhood-specific” information based on Twitter posts. The service provides four types of hyperlocal content: (1) “active events” (current events in the locality by using a statistical event detector that identifies and groups popular features in tweets), (2) “top topics” (most used terms and phrases from recent tweets using a simple topic modeling method), (3) “popular places” (most popular checked-in/mentioned places using both “template-based” and “learning-based” information extractors), and (4) “active people” (Twitter users mentioned the most, using a ranking scheme on the social graph of users) [Hu et al., 2013].

For the same application domain, Bernstein et al.[Bernstein et al., 2010] propose a

¹⁹LinkedIn.com is a social networking website for people in professional occupations

more focused approach for ordering a user's feed into consistent clusters of topics. This means that the proposed framework clusters tweets in a user's feed into topics which have been discussed explicitly or implicitly. This enables users to browse for subjects which appeal to them. For clustering comments into coherent topics, an algorithm has been created for recognizing topics in short status updates. Evaluating the algorithm reveals that enrichment of text (by calling out to search engines) outperforms other approaches by using simple syntactic conversion.

Furthermore, there have been several works which combine machine-based and adaptive ranking approaches which are not necessarily personalized. In this group, a platform proposed by Diakopoulos et al. and DeChoudhury et al. may be mentioned where they examine how journalists filter and assess the variety of trustworthy tweets found through Twitter by using a human centered design approach [Diakopoulos et al., 2012]. They present a number of computational information cues that are useful and effective for this. They have introduced three types of cues: (1) two classifiers, the first classifier classifies users into three types, "organizations", "journalists", or "ordinary people" [De Choudhury et al., 2012]. The second classifier identifies users that might be eyewitnesses to the event, (2) characteristics of the content which are shared by the sources, and (3) characteristics which refer to the event location. With regard to the second classifier, detecting the presence of eyewitnesses is achieved by using supervised learning with manually labeled training examples which include text features.

2.5.3 Summary

Figure 2.8 provides an overview of single-user assessment and ranking of UGC approaches. It is observed that most of the available single-user assessment and ranking approaches focus on maximizing different values mainly for two application domains: postings in micro-blogging platforms and postings in forums. These approaches can be divided into two groups: "Personalized approaches and "Interactive & Adaptive" approaches. The main difference between these two categories is that Interactive & Adaptive approaches in contrast to personalized approaches do not explicitly or implicitly use a user's previous common actions and content to assess and rank the

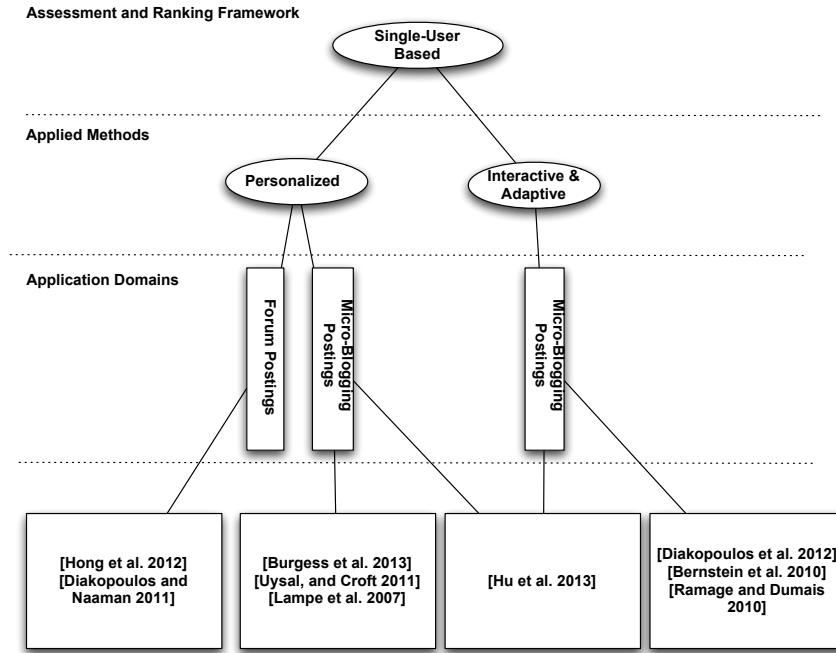


Figure 2.8: Overview of single-user assessment and ranking of UGC approaches

content. However, they provide users with opportunities to interact with the system and explore the ranked content in order to find content to match their requirements. Both categories use approaches which mainly focus on creating interfaces that enable users to more efficiently browse their feed by providing a browsable access to all content in a user’s feed and allowing users to more easily find content related to their interests.

At the backend of these interfaces, there are two types of algorithms: (1) an algorithm which simultaneously leverages the patterns of assessment and ranking settings by a user to minimize the cost of changing settings for other users, and which generally leverage ideas from collaborative filtering and recommender systems [Lampe et al., 2007, Hong et al., 2012, Uysal and Croft, 2011], (2) an algorithm which extracts a set of computational information cues with regard to context and social feed of a user — such as topics [Bernstein et al., 2010] or a set of popular places [Hu et al., 2013] discussed among a set of posts in a user’s feed. Topic modeling based methods and word repetition approaches (such as TF-IDF

or Latent Dirichlet Allocation [Blei et al., 2003]) feature prominently in this space [Ramage et al., 2010]. However, computation of these approaches are costly and noisy, and require too much adjustment to work effectively across a large number of users because users prefer to remove superfluous words from a short posting (such as tweets) to save space. Furthermore, user-generated content have multiple explicit dimensions (such as language tone, physiological aspects, etc.) and, therefore, grouping them based on topic as an exploration facet is a single faceted ranking which does not enable users to rank comments with regard to other potentially useful facets such as subjectivity tone, sentiment polarity, etc. Moreover, providing interpretable descriptions for topic models is rather challenging, and even “ideal” models may not be consistent with viewer preferences [Boyd-Graber et al., 2009].

2.6 Incentivizing high-quality User-generated Content

Voluntary participation and contribution in online social media platforms — contributors may decide to take part or not — is a key factor to be considered when modeling, analyzing, and finally designing mechanisms for assessment and ranking methods [Ghosh, 2012]. It should also be noted that many UGC platforms fail, either immediately or eventually, because of very few contributions. Moreover, having decided to participate does not necessarily mean that contributors will put effort into their contributions [Ghosh, 2012]. This affects the quality of the output they produce. Methods to incentivize contributors need to be developed in order to allocate rewards such as monetary and non-monetary (attention, reputation [Beenen et al., 2004, Huberman et al., 2009], and virtual points) which appear to motivate contributors contrary to what may be expected [Nam et al., 2009, Yang et al., 2011]. Despite there being much need for such methods to incentivize high quality UGC, there are few approaches which focus on development of these approaches.

Ghosh and McAfee [Ghosh and McAfee, 2011] propose a game-theoretic model in the context of diverging attention rewards with high viewership. Strategic contrib-

utors are the focus of the model which is motivated primarily by exposure or viewer attention. The model allows the endogenous determination of both the quality and the number of contributions in a free-entry Nash equilibrium. The importance of making choices to contribute endogenously is underlined because the production of content, and not only incentivizing high quality, is necessary in UGC.

Also, Ghosh and McAfee [Ghosh and McAfee, 2012] explore the design of incentives in environments with endogenous entry for finite rewards. In the context of limited attention rewards in Q&A platforms such as Quora²⁰ or StackOverflow²¹, the choice of which answers to display for each question, the choice whether to display all answers to a particular question, or the choice whether to display only the best ones and suppress some of the weaker contributions remains with the mechanism designer or platform owner. It is demonstrated that when the cost of producing the weakest quality of content is low, then the optimal mechanism is to display all the weakest contributions [Ghosh and McAfee, 2012].

2.7 What Do We Observe, and Where Do We Need Deeper Focus

We present the results of a systematic review of approaches for assessing and ranking UGC with regard to three aspects: “*values which are expected to be maximized*”, “*applied methods*”, and “*application domains*”. We observe that the existing approaches generally adopt one of three frameworks “*Community-Based Assessment and Ranking of UGC*”, which employ machine-based or crowd-based methods, “*Single-User Assessment and Ranking of UGC*”, which employ personalized or interactive & adaptive methods, and finally “*Incentivizing High-quality Content*”. Figure 2.9 shows an overview of ranking and assessment approaches of UGC with regard to adopted frameworks and related utilized methods.

With regard to applied methods, it is observed that most of the proposed assess-

²⁰Quora.com is a question-and-answer website where questions are created, answered, edited and organized by its community of users

²¹StackOverflow.com is a website, the flagship site of the Stack Exchange Network

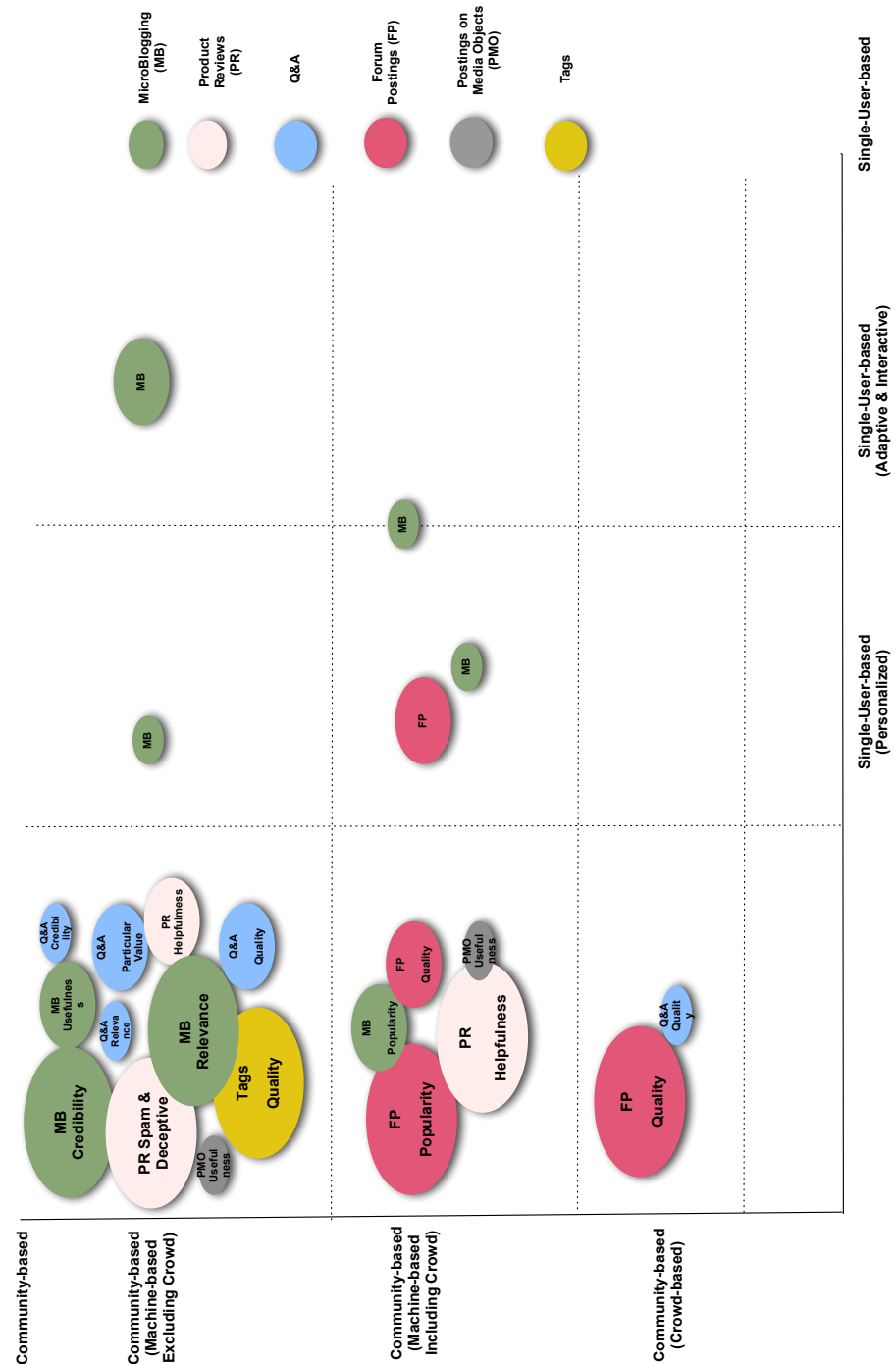


Figure 2.9: Overview of ranking and assessment approaches of UGC. Each bubble represents the amount of work completed for a particular application domain and a value, and the larger a bubble is, the more work has been completed

ment and ranking approaches based on community-based assessment and ranking utilize machine-based methods for assessment of UGC, however, many platforms as a prevalent default approach use the crowd-based approach. Examining machine-based methods more closely reveals that some machine-based assessment approaches include crowd judgments on the content in order to create a ground truth and some completely exclude crowd. On the other hand side, many machine-based approaches exclude crowd for three reasons: (1) different biases of crowd-based approaches such as “imbalance voting”, “winner circle”, “early bird”, etc.[Liu et al., 2007] (2) a lack of an explicit definition of value which may be requested by the crowd to assess some application domains. For example most approaches related to assessment of credibility exclude crowd-based judgments because no platforms or domains have asked the crowd for credibility judgments. (3) human judgments can not be as precise as machine-based judgments in the case of some application domains and values (such as identification of truthful product reviews [Ott et al., 2012]).

Furthermore, it is observed that there are few approaches, which aim to accommodate individual differences in the assessment and ranking of UGC. In other words, there is less consideration of the personalized definition of the value of the individual user and most of the available approaches rely on particular sources of ground truth and do not enable users to make personal assessments of a particular value. For example, most of the work on identification of helpfulness of product reviews creates and develops prediction models based on a set of majority-agreement labeled reviews. However, helpfulness is a subjective concept that can vary for different individual users, and therefore it is important that systems help individuals to make personal assessments of a particular value. Moreover, most of the available assessment and single-user ranking approaches focus on maximizing different values mainly for two application domains: postings in micro-blogging platforms and postings in forums. These works mainly focus on creating interfaces that enable users to more efficiently browse their feed by providing a browsable access to all content in a user’s feed and allowing the user to more adaptively find content related to her interests. At the backend of these interfaces, there are two types of methods: (1) an algorithm which concurrently exploits the patterns of assessment and ranking settings by users to minimize the cost of changing settings for

other users. This method leverages ideas from collaborative filtering and recommender systems [Lampe et al., 2007, Hong et al., 2012, Uysal and Croft, 2011]. (2) an algorithm which extracts a set of computational information cues from a set of content that can be used in the user interface — such as extracting a set of topics [Bernstein et al., 2010]. This means grouping a user’s feed into consistent clusters of related concepts. However, these approaches are sometimes considered to be computationally costly, noisy, and require too much adjustments to work effectively across a wide range of users, — due to the fact that users for saving space remove duplicate words from a short posting. Therefore, alternative approaches which take into consideration the semantic of the content or leverage the users’ social networks for providing high quality groups and subsequent rankings are required [Burgess et al., 2013].

With regard to different values which are expected to be maximized, some features have high impact for assessment of a particular value based on our feature analysis. Therefore, for maximizing some values, systems should take into consideration an easier way to build influential features at the design phase. For example, when maximizing value related to usefulness for comments on online media objects (such as YouTube videos), the system should encourage users and provide them with the opportunity to define references for enriching semantically the text of comments [Momeni et al., 2013a]. In addition, value related to credibility should take authors’ profile pages into consideration [Morris et al., 2012]. Also, many approaches which aim to maximize quality generally apply a crowd-based method. Beside, it is observed that many approaches related to assessment of the relevancy of UGC employ unsupervised learning approaches due to the fact that relevancy is influenced by textual features and, therefore, applying unsupervised text clustering methods is effective for maximizing this value. Many approaches which aim to maximize helpfulness are mainly discussed in the domain of the product review and mainly use crowd-judgments as the ground-truth to build their prediction model. However, the use of crowd for this value is a matter which provokes discussion. Similar to helpfulness, spam and deception are also mainly discussed in the domain of the product review and how they differ in that they mainly exclude crowd-judgments. Approaches which are principally related to the assessment of popularity develop

their identification and prediction models based on votes and ratings of crowd (in the case of Tweeter, re-tweet).

With regard to application domains, a more detailed examination leads to the discovery of many proposed machine-based assessment approaches in the Q&A domain which utilize semi-supervised learning approaches such as co-training or mutually reinforcing approaches. This is due to the high the interconnectedness and interdependency between three sets of entities in Q&A (questions, answers and authors). In addition, most of the available approaches focus on maximizing different values for micro-blogging platforms. This may be due to the very simple and structured characteristics of these platforms. Yet, there are fewer approaches to maximize important values for many application domains such as UGC on online media sharing platforms as an application domain.

Based on these observations our recommendations for the user interface and system designers are: the system which supports the contributions of users should provide an explicit definition of values which are expected to be precisely judged and assessed by crowd. Also, with regard to the high impact of some content and context features for maximizing some values, systems should take into consideration an easier way to build influential features in the design phase. For example, when maximizing value related to usefulness, the system should encourage users and afford them the opportunity to semantically enrich their posts. In addition, value related to credibility should take authors' profile pages into account.

Based on these observations there are a number of challenges which should be taken into consideration for further work. They are as follows:

- How can the conceptual gap between crowd-based and machine-based approaches for optimizing assessment and ranking of the UGC be bridged? This challenge triggers many technical challenges which include: how can we develop algorithms and methods for preventing biases of the crowd, how can we take advantage of semi-supervised learning such as active learning for efficient integration of the crowd into machine-based approaches, or how can we utilize crowd to optimize the process of labeling large amounts of unlabeled UGC and improve the accuracy of hard machine-based judgments?

- How can we help people make personal assessments of a particular value rather than rely on particular sources as authorities for ground truth or minimize the amount of controversial assertions of value among users?
- How can advancement of game-theoretic foundations help incentivize high-quality UGC? There are a number of directions defined by Ghosh [Ghosh, 2012] for the development of the game-theoretic approaches. For example: multi-dimensional model of quality is a more realistic representation of the value of a single contribution. Users at various times after monitoring the existing set of contributions from other users influence their decisions about their own contributed content. Therefore, for a more accurate model, the temporal aspect of UGC may be taken into consideration (sequential model may be better suited to many UGC environments) [Ghosh, 2012].

Chapter 3

Experiments and Datasets

3.1 Introduction

This chapter gives an overview of different experiments carried out for identification of the characteristics of useful comments and creation of usefulness model. The discussion and results of this chapter were published in several conferences and in a journal article [Momeni et al., 2013a, Momeni et al., 2013b, Momeni et al., 2014b, Momeni and Sageder, 2013]. The goal of the work reported in this section is to provide *automated* support for the curation of useful user-generated comments for use as descriptive annotations for digital media objects. To this end, we follow these steps:

1. *Identification of the characteristics of useful comments:* we study two types of media objects — images and videos — from two popular social media platforms — Flickr Commons¹ and YouTube respectively, and collect users’ and experts’ usefulness judgements (by using a crowd-sourcing approach) to identify the usefulness of comments gathered. We then identify technical features that can

¹“The key goals of The Commons on Flickr are to firstly show users hidden treasures in the world’s public photography archives, and, secondly, to show how users’ input and knowledge can help make these collections even richer. Users are invited to help describe the photographs they discover in The Commons on Flickr, either by adding tags or leaving comments.” www.flickr.com/commons

be derived from textual content and the author’s context and characterize the usefulness of a comment.

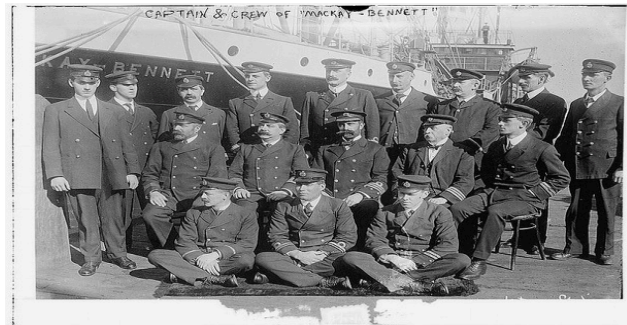
2. *Providing an automated method for identifying potentially useful comments.* We apply the technical features in a series of experiments to build a classifier that can automatically identify the usefulness of comments. Furthermore, we investigate to what extent certain topics of media objects play a role with regard to usefulness classification.
3. *Study the correlation between the commenting culture of a platform with usefulness prediction.* We investigate to what extent the commenting culture of a platform plays a role with regard to usefulness classification.
4. *Study important factors for estimating the prevalence of useful comments.* We adapt an existing model of prevalence detection [Ott et al., 2012] that uses the learned usefulness classifier to investigate patterns in the commenting culture across social media platforms and different dimensions (entity type, time period, and polarization) of topics of media objects.

We investigate usefulness from the users’ perspective, defining a comment as **USEFUL** if it provides descriptive information about the object beyond the usually very short title accompanying it. With this definition in hand, we employ crowd-sourcing techniques to create a gold standard data set of **USEFUL** and **NON-USEFUL** comments and propose the use of standard supervised machine learning techniques to develop a “usefulness” classifier that distinguishes useful from non-useful user-generated comments. We consider over thirty features for the classifier including features for readability, informativeness/novelty, syntactic traits, named entity presence, sentiment, topical traits of the text, and features that describe the author’s posting and social media behavior.

The following examples show some example comments judged as **USEFUL** or **NON-USEFUL** by human coders within our experiments:



Flickr photo - Dr. F.A. Cook



Flickr photo - Capt. and crew of MACKAY-BENNETT

Figure 3.1: Examples of photos of the Library of Congress on Flickr Commons

- **USEFUL: Flickr photo - Dr. F.A. Cook²** (see Figure 3.1 left side). “This must be Dr. Frederick A. Cook (1865-1940), the American explorer who claimed to have reached the North Pole in 1908, before Robert Peary. The controversy over his claim continues. Not only does he have a Wikipedia article, but there are websites dedicated both to disdaining him and to celebrating him. Old controversies never die; they just go on the Internet.”
- **NON-USEFUL: Flickr photo - Capt. and crew of MACKAY-BENNETT³** (see Figure 3.1 right side). “My great grandfather was an engineer at that time. I’d love to get a list of the names in that photo.”
- **USEFUL: YouTube video - Lady diana interview before wedding⁴**. “She had JUST turned 20 years old when they married-in fact it had been less than a month since her 20th birthday. She wasn’t anything more than a teenager. So tell me- how good were you at judging character at that age eh?”
- **NON-USEFUL: YouTube video- World War I: Battle Of Verdun⁵**. “Rich people get their poor people to fight the other rich people’s poor people. And the[n] we do it all over again. Humanity is truly retarded.”

²http://www.flickr.com/photos/library_of_congress/2850357813/comment72157607279573241

³http://www.flickr.com/photos/library_of_congress/2536790306/comment72157629444651496

⁴<http://www.youtube.com/watch?v=Yka3M4uvUyo>

⁵<http://www.youtube.com/watch?v=d2qamDMs-3g>

Our findings can be summarized as follows: first, we find that our trained classifier identifies useful comments for Flickr photos with high reliability (precision (P) of 0.87 and recall (R) of 0.90) and which statistically significantly outperform a strong baseline (P65, R80). However, the identification of useful comments on YouTube proves to be more difficult (P65, R83). Again the classifier statistically significantly outperforms the baseline (P55, R70).

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

Analysis of the top-ranked features of the classifier indicates that semantic and topic-based features are very important for accurate classification for both Flickr and YouTube, especially for those that capture subjective tone, sentiment polarity and the existence of named entities. In particular, comments that mention named entities are more likely to be considered *USEFUL*; those that express the emotional and affective processes of the author are more likely to be considered *NON-USEFUL*. Similarly, terms indicating *INSIGHT* (e.g., think, know, consider) are associated with *USEFULness* while those indicating *CERTAINTY* (e.g., always, never) are associated with *NON- USEFUL* comments.

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

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

3.2 Features Engineering

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

In the rest of this section, we provide an overview of the different features to characterize each comment that we analyze to estimate the usefulness of a comment. Inspired by the cases we found, all these features are aligned with our assumption of characteristics of useful comments. Although we introduce these features by inspiration we got from the Flickr and the YouTube platforms, most of them are quite generic and can also be applied to other platforms. In Table 3.1, we list each feature along with a short description. According to the study we made in Chapter 2, we group these potentially important features into three different groups.

Table 3.1: Overview of Features

Features	Short Description
Text Statistics and Syntactic Features (TS)	
<i>Readability</i>	measures how difficult the comment is to parse using the Gunning fog index [Gunning, 1952]
<i>Informativeness</i>	measures the novelty of terms, t , of a comment, c , compared to other comments on the same object, calculated using: $\Sigma_{t \in c} tfidf(t, c)$

<i>Punctuation Mark</i>	counts the number of punctuation marks
<i>Text Statistics</i>	measures aggregate statistics extracted from the text #Words, #Verbs, #Adverb, WPS (average length of sentences)
<i>Linkage Variety</i>	counts the number of unique hyperlinks in a comment
Semantic and Topical Features (ST)	
<i>Named Entities</i>	counts the number of named entities that are mentioned in a comment
<i>NE Types Variety</i>	counts distinct types of named entities (such as person, place, date, etc.) that are mentioned in a comment
<i>Topical Conformity</i>	measures the distance between the topics of a comment and the topics belonging to other comments on the same object. We use the Jensen-Shannon (JS) divergence to measure the topic distribution distance of all comments on an object (A) compared to the comment's topic distribution (C). $D_{JS} = \frac{1}{2}(D_{KL}(C \parallel A) + (D_{KL}(A \parallel C))$ and KL divergence is calculated as: $D_{KL}(C \parallel A) = \sum_i C(i) \log \frac{C(i)}{A(i)}$.
<i>Sentiment Polarity</i>	measures the sentiment/polarity of a comment as: $SenPolarity = \frac{PositiveScore + NegativeScore}{\#Words}$ We use LIWC for identifying positive and negative scores.
<i>Subjectivity Tone</i>	measures the subjectivity degree of a comment. We use Subjectivity Lexicon [Wilson et al., 2005] to calculate subjectivity
<i>User Topic Entropy</i>	measures the topical focus of an author via the entropy of topic distributions of the author. We define entropy of topic distribution of all comments authored by an author, a_i as: $H(a_i) = -\sum_{j=1}^n p(t_{i,j}) \log p(t_{i,j})$, where t is a topic and n is #topics.
<i>Psychological & Social Characteristics of the Content</i>	identifies psychological dimensions: Leisure, Anger, Family, Friends, Humans, Anxiety, Sadness, Sexuality, Home, Religion, Relativity, Affective Process, and Self-reference scores [Tausczik and Pennebaker, 2010]
User and Social Features (US)	
<i>User Linkage Behavior</i>	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.
<i>User Conversational Behavior</i>	counts comments that contain a @reply
<i>User Activity</i>	measures different activities completed by a user: #Comments (counts the number of comments authored by the user), #Favorite Objects (counts the number of media objects selected as favorite by the user), #UploadedObjects (counts the number of media objects uploaded by the user).
<i>User Social Relation</i>	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)

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

- *Text Statistics* – The aggregate statistics that can be extracted from the comments may also be good indicators for the usefulness of the comments. For instance, the longer comments are more likely to be useful because they have more space for the information and take longer time to be written. A higher number of nouns may indicate that the comment contains knowledge from different aspects. Based on these assumptions, we use the aggregate statistics extracted from the text such as number of words (#WC), number of verbs, number of adverbs, and the average length of sentences (WPS), etc. We collect statistics based on the POS tags to create a set of features such as percentages of verbs, adverbs, etc. We use the LingPipe toolkit⁶ to obtain the relative POS taggers. We hypothesize that comments containing a higher number of words are likely to be useful [Kim et al., 2006b, Ghose and Ipeirotis, 2011].
- *Linkage Variety* – The number of hyperlinks in a comment. The comments written by either experts or users with high relevant knowledge may tend to include the hyperlinks to external credible resources to support their text. Therefore, we hypothesize that the more links are contained in a comment, the more likely it is to be useful [Castillo et al., 2011]. The example below shows a comment with high Linkage Variety, judged as useful by the coders:

“There were 2 different Frances GALLWEY in Tramore. Here is the wife of William GALLWEY. 1901 census, 26 Circus, Bath, Somerset Phyllis DAVIES, Head, Widow, 88, Living on own means, born in Devon, Ugborough Frances K GALLWEY, Daughter, Married, 47, Living on own means, born in Yorkshire, Adlingfleet Jannette P GALLWEY, Granddaughter, Single, 17, Living on own means, born in Ireland plus 5 female servants, all born in Somerset. www.freebmd.org.uk Marriage, March quarter 1883, Bath William Joseph GALLWEY and Frances Kate T DAVIES thepeerage.com/p39134.htm Frances Kate Trelawner DAVIES was the daughter of Reverend Edward William Lewis DAVIES. She married William Joseph GALLWEY, son of Henry Gallwey and Maria Walsh, on 25 January 1883. She died on 29 March 1938. Ireland, Civil Registration Indexes, 1845-1958 Frances K T GALLWEY died in Waterford

⁶<http://alias-i.com/lingpipe/>

*district, 1938.*⁷

- *Informativeness* – This feature measures the novelty of terms used in the comment compared to other comments on the same object. Practically, we use the sum of the TF-IDF⁸ measure to calculate this feature:

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

Here, t is a term used in the comment denoted by c and $tfidf$ is a function, which calculates TF-IDF scores. The higher usage of novel terms in the comment may indicate that it brings more useful information. For that reason, we assume that comments with higher informativeness score are more informative and, therefore, they are likely to be useful [Wagner et al., 2012b].

- *Punctuation Mark* – The number of punctuation marks in the comment. Given that the emotion and a series of meaningless punctuations are frequently seen in comments that are not useful. Therefore, we assume that the number of punctuation marks may have impact on the usefulness of the comments.
- *Readability* – measures how difficult the comment is to parse by using the Gunning fog index [Gunning, 1952]. We assume that comments with a higher readability score are likely to be useful, because they are easier to parse for humans. The example below shows a comment with high readability score, judged as useful by the coders:

*“After being the Boxing Champion of the World, Jimmy Clabby is said to have squandered over \$500,000 in earnings, and was found dead of starvation in Calumet City during the Great Depression.”*⁹

Semantic and Topical Features (ST) Besides superficial identifications, we may get more insights of a comment by checking the semantics. The semantic infor-

⁷Flickr photo - April 15, 1901 <http://www.flickr.com/photos/nlireland/6933777014/comment72157629836757055>

⁸term frequency-inverse document frequency

⁹Flickr photo - Jimmy Clabby. Boxing http://www.flickr.com/photos/library_of_congress/2163449292/comment72157603820313375

mation characterizing from different aspects may have various impact on the likelihood of a comment being useful regardless of its text structure. Furthermore, this group includes standard topical model features, which measure the topical concentration of the author of a comment and the topical distance of a comment compared to other comments made on the same object. Specifically, we analyze the following features:

- *Named Entities* – The number of named entities (NE) that are mentioned in a comment may give evidence on the usefulness. A comment with higher number of entities conveys more concepts that are known to the public. In practice, we use GATE toolkit¹⁰ for the NE related features in this group. We hypothesize that the more entities are identified, the more likely the comment is to be useful. The example below shows a comment with high number of name entities, which is judged as useful by the annotators:

“[Claire L. Runkel (1890 – 1936) and Oscar F. Grab (1886 – 1958) were married on March 23, 1915, at the Ritz Carlton Hotel in New York City. Claire was the daughter of Herman Runkel (1853 – 1918) and Victoria Rebecca Runkel (nee Lopez) (1859-1927), of 150 W. 79th Street in New York City. Mr. Runkel was of the firm Runkel Brothers, chocolate manufacturers. Oscar F. Grab , born Oskar Grab, was an Austro-Hungarian immigrant, United States citizen, and fashion executive. He was a saloon passenger aboard Lusitania who saw the torpedo impact the ship on May 7, 1915.. He saw lifeboats upset on the starboard side and jumped into the water instead of taking a chance in the lifeboats. He was rescued and survived the Lusitania disaster. His wife was not traveling with him. Oscar and Claire moved in with Claire’s parents that October. The couple had two children, Victoria, born in 1916, and Donald born in 1923. Claire also authored a book, By 1928, Oscar’s fashion company, O. F. Grab Company, was a million-dollar business that had branches in France and Belgium and was employing 250 people....]”¹¹

¹⁰<http://gate.ac.uk>

¹¹Flickr photo - (Clara Runkel) Mrs. Oscar F. Grab http://www.flickr.com/photos/library_of_congress/6851810917/comment72157629260546153

- *NE Types Variety* – The number of distinct types of named entities (such as person, place, date, etc.) that are mentioned in a comment. More types of entities mentioned in a comment may indicate that the object is introduced from different aspects. Therefore, we hypothesize that a comment is more likely to be useful if the entities contained in it are more diverse in terms of their types. The previous example also shows the comments with high NE Types Variety.
- *Subjectivity Tone* – The fact or related background knowledge on an object tends to be described in an objective tone. So we assume the subjectivity tone of a comment may impact the usefulness of the comment. By leveraging Subjectivity Lexicon [Wilson et al., 2005], we can calculate the subjectivity of a comment. This enables us to construct the feature of *Subjectivity Tone* with the hypothesis, that a comment with objectivity tone is more likely to be useful. [Ghose and Ipeirotis, 2011]. The example below shows a comment with high objectivity tone, which is judged as useful by the coders:

“Yes, this is the British pavilion by sir Edwin Lutyens.”¹²

- *Psychological & Social Characteristics of the Content* – We can extract psychological and social characteristics of content from the contents by using LIWC [Tausczik and Pennebaker, 2010] for analyzing psychological characteristics. This can give us indicators in various dimensions, including leisure, anger, family, friends, humans, anxiety, sadness, sexuality, home, religion, relativity, affective process, and self-reference. The scores involving authors’ mood, which may be represented by the scores of anger, sadness, may have impact on the usefulness of the comments [Choudhury et al., 2012]. We can suspect that a comment with high score in anger might be written when the author was in a bad mood, therefore is likely to be biased. The example below shows a comment with high anger score, which is judged as non-useful by the coders:

“These pictures are incredible to see especially after reading, The Devil in the

¹²Flickr photo - Paris Exposition: Hungarian Pavilion, Paris, France, 1900 http://www.flickr.com/photos/brooklyn_museum/2486821878/comment72157613666119960

*White City: Murder, Magic, and Madness at the Fair that Changed America.*¹³

- *Topical Conformity* – This feature measures the distance between the topics of a comment and the topics detected in other comments on the same object. An LDA model (Latent Dirichlet Allocation [Blei et al., 2003]), was trained to handle features that depend on topic models. To train the LDA model we aggregated all the comments on objects in our database into an artificial document to infer topic distribution and chose the following 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:

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

The high topical conformity means the comment is closely related to the core message conveyed in the artificial document, and therefore is probably the characteristic of useful comments. For this reason, we hypothesize that the higher the topical conformity we find for a comment the more likely it is to be useful [Weinberger et al., 2008, Wagner et al., 2012b].

- *User Topic Entropy* – The topical focus of an author measured by the entropy of topic distributions of a user may indicate whether she is focusing on some certain topics. This feature can be inferred via the whole set of comments she authored. To handle this feature, we again trained an LDA model [Blei et al., 2003]. For this purpose, we aggregated all the comments authored by each user in our database into one artificial user document to infer topic distribution by her and we chose the following hyper-parameters:

¹³Columbian Exposition: Ferris Wheel, Chicago, United States, 1893. http://www.flickr.com/photos/brooklyn_museum/2784217831/comment72157623500767569

$\alpha = 50/T$, $\beta = 0.01$ and $T = 1,000$. Given the inferred distance topic distribution of each user, we define entropy of topic distribution of all comments authored by an author, a_i as:

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

Here, t is a topic and n is a number of topics. We assume the topical focus of users has influence on the usefulness of their comments.

- *Sentiment Polarity* – Previously, researchers found that the sentiment polarity has an impact on the usefulness of the comments [Ghose and Ipeirotis, 2011, Castillo et al., 2011]. We calculate the sentiment polarity as:

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

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

*“Martins my namesake was great i will be great also. Hahahahaha!!! Interesting.”*¹⁴

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

¹⁴YouTube Video – Martin Luther King, Jr. - Mini Bio http://www.youtube.com/watch?v=3ank52Zi_S0

- *User Linkage Behavior* – The number of unique hyperlinks posted by a user. A high usage of linkage indicates that the author has the behavior of including hyperlinks. As mentioned above, using a hyperlink may support the comment. Here, we evaluate this usage by users. Therefore, we assume that the comments by users that use other resources as references are more likely to be useful.
- *User Conversational Behavior* – On social media platforms, users can interact with each other by writing a comment containing an @reply. The reply messages are frequently found to be questions to previous comments, simple answers to it, or even chat messages. Therefore, we assume that users that write comments to converse with other users are less likely to write useful comments.
- *User Activity* – We can measure the activities completed by a user from different aspects, e.g. the number of comments authored by the user, the number of media objects uploaded by the user, and the number of media objects marked as favorite by the user. The higher these indicators are the more active the user is on the platform. Inspired by [Diakopoulos et al., 2012, Castillo et al., 2011], we construct these features and hypothesize that the more active the user is, the more likely the comments authored by her are seen as useful.
- *User Social Relation* – We measure the social relation of an author by two metrics: the number of contacts that she has and the prestige score measured by the number of the influential contacts (such as Flickr Commons members) in the contact list of the user. We assume that users with a higher number of social interactions are more likely to write useful comments [Lu et al., 2010].

3.3 Data Acquisition

In this section we describe how we collect usefulness judgements for characterizing useful comments. We achieve this by building a dataset from real world comments harvested from Flickr Commons and Youtube which provide free-text comments on media objects (video and photo) from a variety of people with different backgrounds

and intentions, by extracting those comments that have attracted a response by experts of cultural institutes, and finally by using a crowd sourcing approach, setting up a user study, and requesting people to state if they consider that a certain set of comments could be useful for them. Finally in order to show how users’ perception of usefulness is similar to experts’ perception, we compare the characteristics of useful user-judged and expert-judged comments.

3.3.1 List of Topics

In order to analyze the correlation between usefulness and different topics of media objects (topic, time period, etc.), we first selected three types of topics: *event*, *person*, and *place*. Second, we used the history timeline of the 20th century provided by About.com to identify topics associated with the selected topics from each decade of the 20th century. The resulting topics included, among others, 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.

3.3.2 Datasets

Dataset1: we crawled comments written on photos of six different cultural institutes on Flickr Commons. We searched Flickr Commons for photo-sets of each topic (when available) and selected photo-sets which have the highest number of comments on their photos. In one of the Library of congress photo-sets (News in 1910), it is worth mentioning that many of the photos are of persons. Accordingly, photos which show only a photo of a person are separated by us from other photos which belonged to topics related to event, according to their titles. Also, in order to train a classifier and analyze users’ features, we crawl all profile information of all users who wrote comments. Table 3.2 shows the summary statistics of the dataset.

Datasets2: we compiled a dataset from comments harvested from YouTube, searched YouTube for videos of each topic (when available), selected those with the highest number of views and comments (at least 100), and crawled 91,778 comments (the first 1,000 for each topic) written for 310 different videos. For each comment we

Table 3.2: Summary statistics of dataset crawled from Flickr Commons

Photoset	Topic Type	Comments	Objects	Users
Library of Congress	Person, Event	27,603	9,029	4,343
Brooklyn Museum	Place	2,178	251	1,687
National Library of Ireland	Event, Person	1,740	135	470
New York Public Library	Place	251	98	151
National Gallery of Scotland	Place	257	32	201
NASA Collections	Person	103	28	82

Table 3.3: Summary statistics for datasets

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

crawled all the available profile information for the author. As a result of access to some of user’s profile fields being forbidden and limitation of crawling a maximum of 1,000 comments when using the YouTube public API key, we were therefore not able to build all the mentioned features for YouTube dataset (such as Informativeness or Topical Conformity and some user related features such number of contact, prestige score, number of favorite objects) in the feature engineering phase.

In total for Flickr we crawled 33,273 comments written on 11,102 photos. For YouTube we crawled 91,778 comments (the first 1,000 for each topic) written for 310 different videos. (Distribution of the comments across different topics is shown in Table 3.3.) As a result, we obtained comparable datasets from YouTube and Flickr for topics involving events, people, and places across different time periods starting in 1900.

3.3.3 Collecting User Judgements for Defining Usefulness

We randomly selected 3,500 comments from Flickr and 5,000 from YouTube in order to code manually with respect to usefulness. (As will be seen below, more comments were required from YouTube due to the low rate of useful comments.) See Table 3.4 for results of manual coding.

Table 3.4: Manual coding results across platforms. Agreement scores are assessed based on Mean Fleiss’ Kappa scores.

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

Coders were found via the CrowdFlower.com crowd-sourcing platform which distributed our task across different channels, such as Mechanical Turk or getPaid. Coders were asked to assist us to define useful comments, by showing each coder a comment and links to the related media object (Flickr photo or YouTube video). That the work by coders meets with high quality, we asked coders to answer four objective questions for each comment. The answers to the first three questions can be computed automatically, and a fourth question addressed the usefulness of the comment. The first and second questions for both platforms were semantically the same but asked in two different ways. Inconsistency in answering the first two questions gives us the chance to exclude randomly selected answers. The first two questions for the Flickr user study are: 1- “How many Web links does the comment contain?” and 2- “Does the comment contain Web links”? The first two questions for the YouTube user study are: 1- “Is the length of the video short or long?” (more than two minutes is long, less than two minutes is short) and 2- “Does the comment contain Web links”? The third question was the following: 3- “How long is the length of the video?”. This question required writing a text-based answer, offering an additional chance to exclude data from non-serious coders. The main question (the fourth question) for the task was the following: 4- “Compared to the description provided by the uploader of the media object (located below the video or photo), is this comment useful for you to learn more about the content of the media object (video or photo)?”. For each comment we collected three judgements.

In order to prepare a training-set for developing a usefulness classifier, first, we select 1,000 user-judged useful comments with high agreements on being useful and 1,000 comments with high agreements on being non-useful from our labeled data. Second, we assess the mean values and standard deviations of each feature, as shown in Table 3.5. As expected, the average semantic and topical-based scores for comments

Table 3.5: 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

Features	Flickr				YouTube			
	Mean-U	STD-U	Mean-N	STD-N	Mean-U	STD-U	Mean-N	STD-N
Text Statistics and Syntactic Features (TL)								
<i>Readability</i>	06.05	04.07	05.70	03.54	<u>09.12</u>	<u>07.87</u>	<u>05.46</u>	<u>05.38</u>
<i>#Punctuation Marks</i>	77.76	131.4	77.10	214.7	25.02	28.63	32.06	44.17
<i>#WC</i>	41.70	49.41	09.32	12.52	41.17	31.77	19.82	19.79
<i>#WPS</i>	<u>15.63</u>	<u>10.99</u>	<u>06.36</u>	<u>06.50</u>	<u>20.47</u>	<u>17.99</u>	<u>12.51</u>	<u>11.79</u>
<i>#Verb</i>	09.06	08.61	09.05	11.38	13.61	06.99	14.34	11.02
<i>#Adverb</i>	02.91	04.81	05.10	10.30	04.54	05.52	04.80	07.37
<i>Linkage Variety</i>	<u>01.72</u>	<u>01.82</u>	<u>0.521</u>	<u>05.92</u>	–	–	–	–
<i>Informativeness</i>	<u>14.50</u>	<u>21.91</u>	<u>05.02</u>	<u>06.37</u>	–	–	–	–
Semantic and Topical Features (ST)								
<i>#Name Entities</i>	<u>03.62</u>	<u>05.33</u>	<u>0.466</u>	<u>0.956</u>	<u>02.44</u>	<u>02.77</u>	<u>01.07</u>	<u>01.67</u>
<i>NE Types Variety</i>	<u>01.39</u>	<u>01.07</u>	<u>00.36</u>	<u>00.58</u>	<u>01.10</u>	<u>00.83</u>	<u>0.639</u>	<u>0.704</u>
<i>Topical Conformity</i>	01.34	01.67	01.07	01.10	–	–	–	–
<i>Sentiment Polarity</i>	<u>01.62</u>	<u>03.75</u>	<u>29.26</u>	<u>32.77</u>	<u>06.59</u>	<u>09.01</u>	<u>10.44</u>	<u>15.28</u>
<i>Subjectivity Tone</i>	<u>0.151</u>	<u>0.160</u>	<u>0.910</u>	<u>0.750</u>	<u>0.187</u>	<u>0.122</u>	<u>0.296</u>	<u>0.265</u>
<i>Sadness</i>	0.190	0.880	0.160	0.940	0.411	0.129	0.562	04.09
<i>Insight</i>	<u>0.150</u>	<u>01.56</u>	<u>0.096</u>	<u>0.810</u>	01.48	02.35	01.66	03.90
<i>Anger</i>	0.369	01.74	0.197	01.80	<u>01.91</u>	<u>05.92</u>	<u>02.41</u>	<u>07.29</u>
<i>Family</i>	0.460	01.63	0.126	01.40	0.359	01.42	0.329	01.74
<i>Friends</i>	<u>0.060</u>	<u>0.950</u>	<u>0.130</u>	<u>02.98</u>	0.049	0.497	0.087	01.10
<i>Humans</i>	0.590	01.93	0.840	03.64	01.33	03.49	01.26	03.88
<i>Health & Body</i>	<u>0.790</u>	<u>02.41</u>	<u>01.93</u>	<u>07.02</u>	<u>01.29</u>	<u>03.52</u>	<u>02.28</u>	<u>06.65</u>
<i>Sexual</i>	<u>0.065</u>	<u>1.086</u>	<u>0.970</u>	<u>05.10</u>	<u>0.356</u>	<u>0.528</u>	<u>01.06</u>	<u>05.00</u>
<i>Religion</i>	<u>0.409</u>	<u>02.86</u>	<u>0.103</u>	<u>01.21</u>	0.404	0.30	0.61	03.50
<i>Leisure</i>	01.30	02.99	0.460	02.51	01.29	02.75	01.57	05.60
<i>Swear</i>	<u>0.058</u>	<u>0.087</u>	<u>0.198</u>	<u>0.682</u>	<u>0.216</u>	<u>01.44</u>	<u>01.33</u>	<u>06.12</u>
<i>Home</i>	0.450	01.74	0.180	01.35	0.091	0.515	0.167	01.07
<i>Relativity</i>	12.86	09.18	06.14	09.87	<u>12.61</u>	<u>08.46</u>	<u>10.23</u>	<u>11.07</u>
<i>Certainty</i>	0.616	1.980	1.290	6.750	01.54	02.81	01.97	05.37
<i>Tentative</i>	01.79	03.65	01.21	03.98	02.07	03.22	02.00	04.72
<i>Self-reference</i>	<u>01.02</u>	<u>2.587</u>	<u>02.27</u>	<u>05.42</u>	<u>01.24</u>	<u>02.73</u>	<u>03.08</u>	<u>06.19</u>
<i>User Topic Entropy</i>	04.74	01.67	04.34	02.69	–	–	–	–
User and Social Features (US)								
<i>User Linkage Behavior</i>	<u>758.0</u>	<u>1225</u>	<u>09.93</u>	<u>88.44</u>	–	–	–	–
<i>User Conversational Behavior</i>	<u>0.480</u>	<u>02.35</u>	<u>19.20</u>	<u>33.65</u>	0.522	0.501	0.392	0.488
<i>#UploadedObject</i>	20250	3869	1390	3134	<u>11.17</u>	<u>64.94</u>	<u>05.46</u>	<u>34.48</u>
<i>#FavoriteObject</i>	243.5	220.5	269.1	219.5	–	–	–	–
<i>#Contact</i>	179.1	261.7	204.6	283.6	–	–	–	–
<i>Prestige score</i>	<u>04.96</u>	<u>09.61</u>	<u>01.62</u>	<u>4.274</u>	–	–	–	–

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

3.3.4 Collecting Expert Judgements for Defining Usefulness

With regard to comments written on photos of the Library of Congress (LOC), we notice some of these comments are commented upon by the LOC experts¹⁵. In order to ensure that these comments are useful for LOC, we ask LOC staff members why they comment back. They confirm that commenting back is one indicator of a useful comment: “all Flickr comments are being read by LOC staff. The vast majority of comments is useful, but we only have the resources to comment back when we verify that a suggested change was on target, so that the Flickr users know that their information is making a difference.”. Based on these observations, first we

¹⁵These are user accounts which have the pattern “Name (LOC P&P)” and use the Library of Congress logo

Table 3.6: The comparison of the mean and standard deviation values of each feature between user-judged (U) and expert-judged (E) useful comments.

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

crawl all comments written by LOC staff and containing terms such as “thanks”, “thank you”, etc. Second, in order to find related comments to these comments, we use the crowd sourcing approach and we ask coders to assist us in defining relevant comments. We use CrowdFlower.com which is a crowd-sourcing platform, showing each coder a comment written by LOC staff and links to the related Flickr photo and asking them to find all relevant comments to LOC experts’ comments. In total we gather comments amounting to 2,068, which we presume to be considered useful by experts. It is worth mentioning that LOC experts have not explicitly classified comments as useful and non-useful. This means that comments which in our study are inferred as “non-useful” might be useful for other contexts.

Furthermore, in order to compare characteristics of useful user-judged with useful expert-judged comments we randomly selected 1000 useful expert-judged useful comments and we selected 1,000 useful user-judged comments with high agreements on being useful.

Second, we assess the mean values and standard deviations of each feature for expert-judged comments. Table 3.6 shows in detail these values in comparison with user-judged useful comments. This table shows the mean and standard deviations of almost all features from both datasets are in the same range. This result suggests that the characteristics of user-judged comments are very similar to characteristics of expert-judged useful comments and therefore the non-useful comments (labeled in our study) can be assumed to be non-useful from both perspectives.

3.4 Experiments

In this section, we introduce the process of building the learning-based “usefulness” classifier and evaluate it on the manually coded comments. Given the comments on which the usefulness is estimated, we calculate all the features introduced in Section 3.2 and attempt to build the classifier. The classifier can then be automatically used as an inference method to predict whether a comment is useful or not. We report on the estimation performance by applying different machine learning algorithms. Next, we evaluate the importance of the features that can be interpreted from the

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

3.4.1 Usefulness Classifier

Experimental Setup For training the usefulness classifier, we selected a balanced set of 1,000 `USEFUL` comments and 1,000 `NOT USEFUL` comments from the Flickr data; we selected 400 of each class from the YouTube data. Each of these comments has been judged at least three times, by different coders. Moreover, to ensure the quality of the judgements, the comments may be selected only with majority agreement on usefulness or being not useful. Practically, we have employed two machine learning algorithms, logistic regression (LR) and Naive Bayes (NB), to build the classifiers. Classifiers were trained by using different combinations of the feature groups described in Section 3.2. For evaluation, we focus on four measures: precision (P), recall (R), F1-measure (F1), and area under the Receiver Operator Curve (ROC). Besides the proposed usefulness classifiers, we designed two baseline approaches for comparison purposes:

Baseline 1 predicts usefulness by using the feature of `INFORMATIVENESS`. This feature is demonstrated by Wagner et al. [Wagner et al., 2012b] to be an influential feature for predicting the attention level of a posting in online forums.

Baseline 2 predicts usefulness by using the feature of `SUBJECTIVITY TONE`, which is a particularly strong baseline as a result of our feature analysis study (see Section 3.4.2).

Results of Evaluations of Different Classifiers Table 3.7 provides an overview of both the estimation performances of the two baselines and classifiers trained with different combinations of the feature groups by using two machine learning algorithms. The results demonstrate the effectiveness of using semantic (ST) and user-related (US) features for inferring useful comments.

In particular, for both Flickr and YouTube datasets, the classifiers created by using author and semantic features outperform the models trained with text features (TS)

Table 3.7: Results from the evaluation of classification algorithms with different feature settings (**bold** indicates the top F1 and ROC scores for each dataset)

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

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

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

3.4.2 Influence of Features on Usefulness Classifier

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

Table 3.8: Top-20 features for each platform and related coefficient ranks derived from the Logistic Regression model. Features are ranked based on Information Gain Ratio.

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

To investigate how the features were associated with the usefulness of comments, we examine the coefficients of the best-performing Logistic Regression model (using *ALL* groups of features). Table 3.8 lists the coefficients of 20 features that are highly ranked in terms of Information Gain Ratio (IGR). The features with positive coefficients are positively correlated to the usefulness while the negative coefficients are negatively correlated to the usefulness. Following, we analyze these results and try to validate our hypotheses made in Section 3.2.

Results of Influential Features The top-ranked features from two datasets are both dominated by Semantic and Topical (ST) features. More specifically, coefficient ranks show that comments that express emotional and affective processes of the author (higher *Subjectivity Tone*, *Sentiment Polarity*, *Anger*, *Sadness*, *Swear*, and *Anxiety* scores) are more likely to be inferred as NOT USEFUL. *Subjectivity Tone* is a very good indicator for both platforms. Higher *Subjectivity Tone* has negative impact on the usefulness classifier. Therefore, we have the hypothesis made in Section 3.2 validated. Furthermore, comments with offensive language (higher *Swear* score) are more likely to be inferred as NOT USEFUL. An analysis of the *Swear* and *Anger* scores between different platforms shows that YouTube contains more

offensive language. Therefore, the *Swear* and *Anger* scores for YouTube are more negative than the Flickr swear score. This can be explained by that more frequent emotional comments are posted on YouTube, while on Flickr this is not the case. Besides, the ranks show that comments that have higher number of *Named Entities*, *NE Type Variety*, and *Linkage* scores contain potentially interesting information and are likely to be inferred as *USEFUL*. Therefore, we confirmed the assumption made for Named Entity related features.

We have constructed a series of features with the name of “Psychological & Social Characteristics of the Content” (see Section 3.2) by using LIWC. The usage of terms in LIWC’s **insight** category (such as think, know, consider) shows positive correlation with usefulness on the Flickr dataset. This is in line with the relatively high difference of this feature between *USEFUL* and *NOT USEFUL* comments. Furthermore, terms in LIWC’s **certainty** category (such as always, never) has a negative impact on the model. This might be due to the fact that authors who are assertive and express certainty tend to be seen as more subjective and less analytical. In contrast, using terms in LIWC’s **tentative** category (such as maybe, perhaps, guess) shows that authors make less claims as to the correctness or certainty of their comments and such comments are likely to be determined *USEFUL*.

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

With regard to User & Social (US) features, *User Linkage Behavior* is a good indicator showing that authors may diligently cite references for the information they provide. This increases reliability when inferring such comments as *USEFUL*. Similarly, we note that a higher *Linkage* score has a positive impact on the usefulness inference, which is in line with the correlation of User Linkage Behavior score. Consequently, we can confirm the hypotheses made for these two features. A higher score of *Self-reference* and a higher *User Conversational* score have a negative impact. This suggests that authors who mostly use systems to converse and describe

their personal experiences do not write useful comments. Again, we have validated our thoughts while constructing these two features. Interestingly, a higher *User Topical Entropy* score of authors has a negative impact on the usefulness inference. This indicates that authors with a higher entropy have a lower topical focus and therefore write a comment with a lower level of focus and knowledge about the specific topic. Therefore, their comments are likely to be inferred as NOT USEFUL.

Results of Iteratively Appending Features In order to observe the impact of iteratively appending features on classification performance, we conduct a further experiment to investigate how the performance of the classifiers changes as the top-ranked features are increasingly added for training. In particular, we apply the Logistic Regression algorithm for training - based on its optimum performance during the model selection phase - and trained the classifier using the training split from the first dataset. In Figure 3.2 we can see how the performance of the classifier changes with more and more top ranked features. The result shows the classifier can achieve about 70% and 80% of best performance in terms of F1 and ROC respectively with only one feature. With top 7 features, the trained classifier can already achieve about 90% and 95% of the optimal F1 and ROC respectively. By further adding features ranked lower, we observe similar levels of performance.

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

3.4.3 Influence of Topic on Classification

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

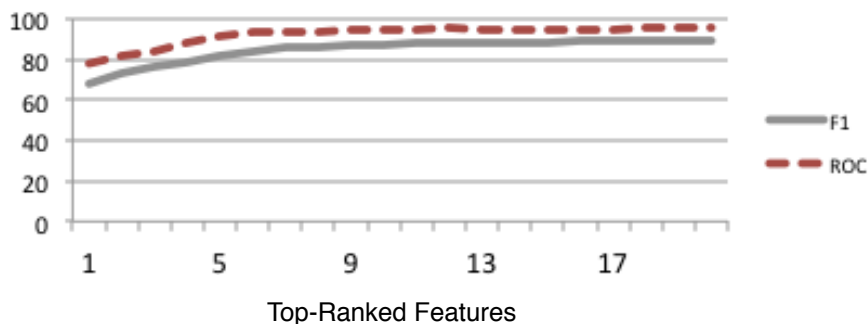


Figure 3.2: Performance results of classification using top-20 features (Results of Iteratively Appending Features)

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

Results from the evaluation of usefulness classifiers for different topics The performance results for type-specific and type-neutral classifiers are given in Table 3.9. We find that, in general, performance is better when the classifier is trained on comments of a single type, i.e., the classifier is type-specific, whereas performance is worse when the type is ignored, i.e., the classifier is type-neutral. We additionally perform three Pearson’s Chi-squared tests between the prediction results of each classifier for each topic. In Table 3.9, “*” indicates a significant difference at a $p < 0.01$ level for some types. We can conclude that it at least makes sense to build a specific model for the object type of Person or Place.

Furthermore, we investigate the importance of features for each topic type of object with regard to usefulness inference. Table 3.10 shows detailed coefficient ranks for different models of three types of topics. Our discussion of the results focuses on the difference between the classifiers derived for each of the topic types. An analysis of the most important features among different type of objects (Person, Place, and Event) shows some differences. The major differences appear among the features related to “Psychological & Social Characteristics of the Content”, but a few

Table 3.9: Results from the evaluation of usefulness classifiers for different object types. *All* is the type-neutral classifier, which is trained on data corresponding to all topic types of objects. “*” indicates a significant difference ($p < 0.01$).

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

differences appear among other semantic and user features. There is no significant difference among text features.

More precisely, coefficient ranks show that comments related to the type of topic, Person and Event, express the author’s emotional and affective processes more. These contribute to a comment being classified as NOT USEFUL. An analysis of the *Subjectivity Tone* among different topics shows that the *Subjectivity Tone* for objects related to Person is higher than for other types. This can be explained by that authors of NOT USEFUL comments tend to use a subjective tone. An analysis of the *Swear* score among different topic types shows that the *Swear* score for the topic type Person is the most negative one. With regard to the objects related to Event, the *Swear* score is more negative than for topics related to place.

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

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

for an object with topic related to Place is NOT USEFUL.

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

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

As shown in Table 3.9, the result demonstrates that different platforms (Flickr and YouTube) lead to performance differences in usefulness classification. For all topic types (Place, Person, and Event), the performance of usefulness classifiers derived from Flickr platform is higher than that from the YouTube platform. Besides the data limitation mentioned before (see Section 3.3), this may also be caused by cultural differences in commenting behaviors.

Furthermore, we investigate each feature by comparing the difference in coefficients in usefulness classifiers built with two platforms. For Flickr, we note a higher *Contact* score does not have a negative impact. However, a *Prestige* score has a positive impact. This indicates that having influential contacts in the contact list is more important than having a higher number of contacts. For YouTube, users with a higher number of uploaded objects are more likely to write useful comments. This does not apply to Flickr. No comparison can be made between YouTube and Flickr on contact related features due to the lack of crawled data from YouTube

The above experimental results indicate that:

- There are a few relatively straightforward features that can be used to infer usefulness of user-generated comments.

Table 3.10: Top-20 features for each platform and related coefficient ranks derived from the Logistic Regression model for each topic. Features are ranked based on Information Gain Ratio (IGR in type-neutral classifier?).

Flickr				YouTube			
Feature	Place	Person	Event	Feature	Place	Person	Event
ST-Subjectivity Tone	-	-	-	ST-Subjectivity Tone	-	-	-
	4.271	6.228	3.406		0.129	2.386	2.002
ST-Sentiment Polarity	-	-	-	ST-#Name Entities	0.049	0.124	0.209
	0.157	0.223	0.647				
ST-NE Types Variety	-	0.113	0.776	ST-Self-reference	-	-0.46	-
	0.138				0.148		0.360
US-User Linkage Behavior	0.046	0.003	0.002	ST-Swear	-	-	-
					0.002	0.571	0.145
ST-#Name Entities	0.203	0.109	0.201	ST-Sentiment Polarity	-	-	-
					0.023	59.73	0.173
ST-Self-reference	-	-	-	ST-NE Types Variety	-	-	0.328
	0.161	0.136	0.177		0.109	0.175	
ST-User Topic Entropy	-	-	-	ST-Anger	-	-	-
	0.112	0.302	0.059		0.188	0.138	0.131
ST-Insight	-	0.081	0.064	ST-Tentative	0.171	0.051	0.120
	0.124						
ST-Swear	-	-	-	US-#UploadedObject	0.015	1.556	0.014
	0.005	90.42	3.363				
TS-Linkage	0.084	3.028	0.610	TL-Future Verb	-	-	-
					0.426	0.182	0.298
AS-User Conversational	-	-	-	ST-Certainty	0.023	-	-
	0.086	0.086	0.066			0.034	0.003
ST-Certainty	0.110	0.042	-	US-User Conversational	-	-	0.083
			0.054		0.154	0.484	
TS-Future Verb	-	-	-	ST-Anxiety	-	-	0.008
	0.071	0.027	0.027		0.216	0.339	
TS-Impersonal-pronoun	-	-	-	TS-Impersonal-pronoun	-	0.041	-
	0.052	0.040	0.042		0.018		0.087
US-Prestige score	0.162	0.005	0.070	ST-Friend	-	-	-
					0.519	0.046	0.011
ST-Religion	0.361	0.322	0.089	ST-Religion	0.046	-	0.021
						0.017	
ST-Sadness	-	-	-	ST-Sadness	0.325	-	0.289
	0.110	0.403	0.038			0.218	
ST-Sexual	-	-	-	ST-Sexual	-	-	-
	1.306	0.812	0.284		0.007	0.175	0.059
ST-Family	-	1.111	-	ST-Home	-	0.692	-
	0.196		0.004		1.760		0.611
ST-Relativity	0.163	-	0.029	ST-Family	-	0.352	0.031
		0.160			0.233		

- An analysis of the important features across different platforms and different object types reveals that when inferring usefulness, the impact of features varies slightly.
- The major differences appear among the psychological and social features (derived from LIWC) of the content. Therefore, a classification model should be trained that takes the topic of media object into account for building type-specific usefulness classifiers with higher accuracy.
- The commenting cultures on different social media platforms are different. Therefore, a classification model should be trained that takes the commenting culture of a platform into account for building the usefulness classifiers.

3.4.5 Prevalence of Useful Comments

This section aims to understand patterns in authors' comments peculiar to a particular commenting culture on different platforms and different dimensions (entity type, time period, and polarization) of topics of media objects. For estimating the prevalence of useful comments we adapt an existing Bayesian Prevalence Model [Ott et al., 2012] that uses the learned usefulness classifiers (see Table 3.9). The Bayesian Prevalence Model estimates the prevalence of useful comments in a set of comments by correcting the output of the noisy usefulness classifiers based on the performance characteristics of the classifiers. In the following section, first, we describe the formal definition and usage of the Bayesian Prevalence Model in our scenario and then we describe our experimental set up for estimating the prevalence of useful comments.

Bayesian Prevalence Model

Given an imperfect usefulness classifier, f , and a set of unlabeled comments, C_U , our goal is to use f to estimate the rate, or prevalence, of useful commenting in C_U . This task is challenging since f can produce both false positive and false negative predictions, and, therefore, cannot be relied on directly. Furthermore, if the

probability of a false positive is different from the probability of a false negative, then the error introduced by f will vary depending on the true rate of useful commenting in C_U .

To address these challenges, we adopt the Bayesian Prevalence Model, introduced by Ott et al. [Ott et al., 2012] to estimate the prevalence of deceptive online reviews, and jointly model our classifier’s false positive and false negative rates, as well as the true rate of useful commenting in C_U . Formally, let us define our classifier, $f : \mathbf{c} \rightarrow y$, as a function mapping a comment, $\mathbf{c} \in \mathbb{R}^{|V|}$, to a usefulness label, $y \in \{0, 1\}$, where $|V|$ corresponds to the number of features. We further define f ’s *sensitivity* (true positive rate), η^* , and *specificity* (true negative rate), θ^* , as:

$$\begin{aligned} \text{sensitivity} &= \eta^* = \Pr(f(\mathbf{c}) = 1 \mid y = 1), \\ \text{specificity} &= \theta^* = \Pr(f(\mathbf{c}) = 0 \mid y = 0). \end{aligned}$$

Then, in order to estimate the true rate of useful commenting in C_U , π^* , we model the process by which f makes its predictions. In particular, we model predictions made by f as a generative process with the following storyline:

- Sample the rate of useful commenting: $\pi^* \sim \text{Beta}(\boldsymbol{\alpha})$
- Sample the classifier’s sensitivity: $\eta^* \sim \text{Beta}(\boldsymbol{\beta})$
- Sample the classifier’s specificity: $\theta^* \sim \text{Beta}(\boldsymbol{\gamma})$
- For each comment, \mathbf{c} , in C_U :
 - Sample the comment’s usefulness: $y \sim \text{Bernoulli}(\pi^*)$
 - Sample the classifier’s prediction:

$$f(\mathbf{c}) \sim \begin{cases} \text{Bernoulli}(\eta^*) & \text{if } y = 1 \\ \text{Bernoulli}(1 - \theta^*) & \text{if } y = 0 \end{cases}$$

Following Ott et al. [Ott et al., 2012], we treat η^* and θ^* as latent variables with prior probabilities, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, set based on the cross-validation results in the previous

section (see Table 3.9). We perform inference for this model with 70,000 iterations of Gibbs sampling, with 20,000 burn-in iterations and a sampling lag of 50. See Ott et al. [Ott et al., 2012] for sampling equations and full derivation details.

Experimental Set Up

We set up three different experiments. First, for exploring the influence of time periods of topics on usefulness prevalence, we create 10 sets of comments related to each decade of the 20th century. Second, to explore the influence of a topic’s polarization on its usefulness prevalence, we create 10 sets of comments from topics with varying degrees of polarization. Third, for exploring the influence of entity types of topics on usefulness prevalence, we create 6 sets related to each platform, that is for each platform one set for each entity type of topic (person, place and event), in total 26 sets. For each set of each experiment we used the trained usefulness classifiers (see Table 3.9) and we predicted the usefulness of each comment and then we instantiated the Bayesian Prevalence Model in order to estimate the realistic rate of the different sets of comments related to the different dimensions of topics.

Influence of Time Periods of Topics on Usefulness Prevalence. In order to observe the effects of the time period of the topics (e.g, year of an event) on the prevalence of useful comments, we explore the prevalence for useful comments among different time related sets of comments, which belong to different time periods (different decades of the 20th century). Our results (shown in Figure 3.3) demonstrate that the temporal dimension of topics has slight influence on the usefulness prevalence. The nearer the time period of a topic is to the present time, the lower the prevalence of useful comments is. This might be due to the fact that topics related to earlier periods are less relevant to present time, therefore authors express less emotion and give more objective information, which may be inferred as useful information.

While these results (Figure 3.3) are useful for an initial analysis, a statistical analysis is required to draw more certain conclusions to demonstrate that the prevalence of useful comments varies with regard to the temporal dimension of topics. Therefore, we apply Pearson’s Chi-squared tests between the prevalence results of each of the

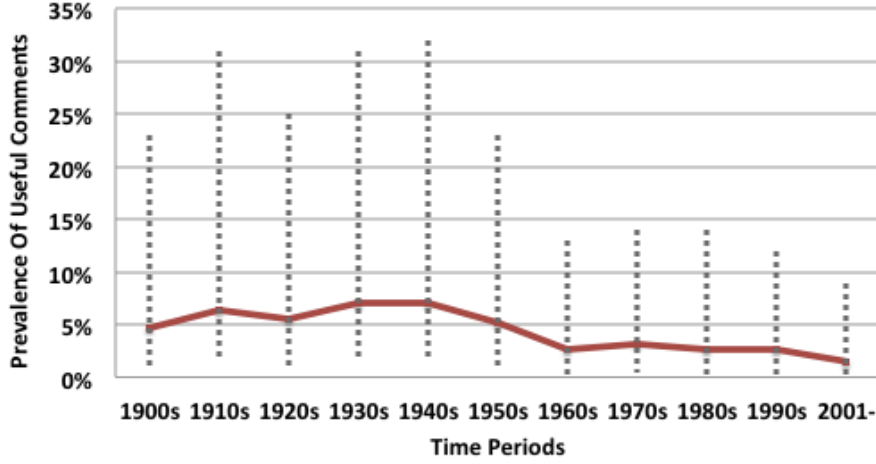


Figure 3.3: Graph of Bayesian estimates of usefulness prevalence versus time periods. Error bars show Bayesian 95% credible intervals.

two different time periods in our dataset. The results of this study (shown in Table 3.11 indicate that there are statistically significant differences between various topics related to different time periods.

Influence of Polarization Degree of Topics on Usefulness Prevalence. Our next experiment explored the relationship between the prevalence of useful comments and the polarization degree of topics of media objects. Following Siersdorfer et al. [Siersdorfer et al., 2010], by “polarizing topic” we mean a topic likely to trigger diverse sentiments and opinions among commenters, such as topics related to a presidential election in contrast to rather “neutral” topics such as “Ford Introduces the Model-T”. In order to assess the polarization degree of topics we leverage the results of an exciting study [Siersdorfer et al., 2010] on the polarization of YouTube videos, which show that polarizing videos tend to trigger more diverse user-rating behaviors on comments and video. For identifying polarizing videos, we compute the difference of video and comments user-ratings. Thus, we compute the difference between the numbers of thumbs up (t_u) and thumbs down (t_d) as: $polarization = 1 - |(t_u - t_d)/(t_u + t_d)|$ for each video in our dataset¹⁶. Using this method our polarization range is between $[0, 1]$. For polarization range we derive 10

¹⁶This experiment was conducted only on a YouTube set and not on a Flickr set, because Flickr photos do not have any rating.

Table 3.11: Results of significant differences between prevalence of usefulness for various topics related to different time periods. The numbers indicate X^2 values of Pearson’s Chi-squared tests between each two time periods. “*” indicates a significant difference with $p < 0.01$). “**” indicates a significant difference with $p < 0.001$). “***” indicates a significant difference with $p < 0.0001$).

	2000-	1990s	1980s	1970s	1960s	1950s	1940s	1930s	1920s	1910s
1900s	50.29***	52.67***	33.67***	39.49***	71.13***	0.022	68.21***	16.98***	0.098	19.53***
1910s	223.9***	199.3***	155.0***	144.2***	235.0***	31.76***	15.77***	0.021	6.796**	–
1920s	134.7***	124.6***	90.14***	87.26***	154.3***	7.875**	48.67***	5.375*	–	–
1930s	190.9***	174.9***	135.2***	128.4***	207.7***	26.78***	15.45***	–	–	–
1940s	586.6***	451.7***	375.0***	315.4***	501.9***	121.6***	–	–	–	–
1950s	85.50***	80.06***	50.99***	51.28***	106.7***	–	–	–	–	–
1960s	8.46**	2.568	12.07***	5.326*	–	–	–	–	–	–
1970s	0	0.745	0.633	–	–	–	–	–	–	–
1980s	1.174	3.565	–	–	–	–	–	–	–	–
1990s	1.179	–	–	–	–	–	–	–	–	–

Table 3.12: Results of significant differences between the prevalence of useful comments between topics with various polarization values. The numbers indicate X^2 values of Pearson’s Chi-squared tests between each two various polarization values. “*” indicates a significant difference with $p < 0.01$). “**” indicates a significant difference with $p < 0.001$). “***” indicates a significant difference with $p < 0.0001$).

	0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9
0.9-1	9.775***	168.2***	1.133*	239.0***	1.81	39.83***	92.55***	57.87***	201.5***
0.8-0.9	467.2***	20.92***	486.2***	54.58***	643.2***	110.4***	95.14***	314.9***	–
0.7-0.8	269.8***	61.92***	226.3***	113.0***	348.2***	16.97***	0	–	–
0.7-0.6	251.1***	59.93***	195.1***	109.7***	305.3***	15.01***	–	–	–
0.5-0.6	123.9***	84.99***	57.71***	139.9***	115.6***	–	–	–	–
0.4-0.5	5.885***	267.4***	12.72***	362.0***	–	–	–	–	–
0.3-0.4	376.1***	4.756***	285.4***	–	–	–	–	–	–
0.2-0.3	28.42***	202.4***	–	–	–	–	–	–	–
0.1-0.2	281.9***	–	–	–	–	–	–	–	–

bins (such as 0-0.1). Comments on videos are assigned to a particular bin depending on the polarization topic of the related video. Then we estimate the prevalence of useful comments for each set related to each bin by using the usefulness classifiers and the Bayesian Prevalence Model.

The result of the relationship between the prevalence of useful comments and the polarization of topics of media objects is shown in Figure 3.4. We find the prevalence of useful comments decreases when the polarization of topics is higher. Furthermore, we inspected the coefficients of a linear regression model between the prevalence of useful comments on each video and polarization degree of the video. The coefficient

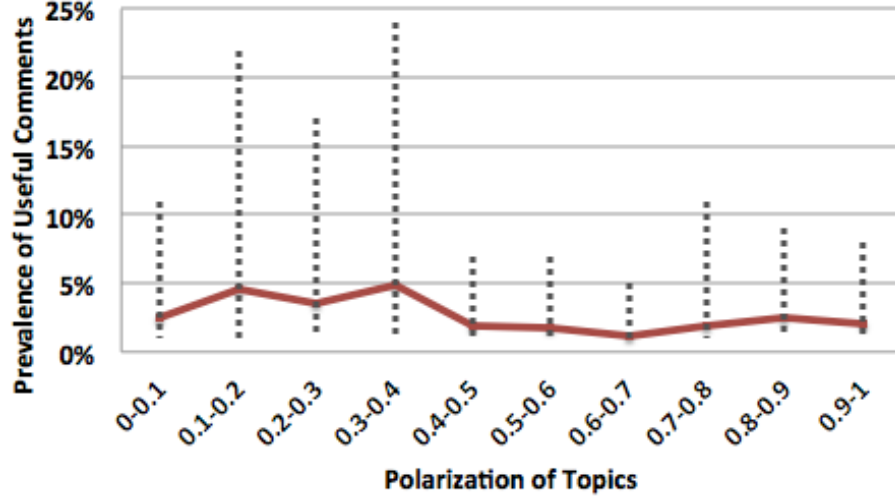


Figure 3.4: Graph of Bayesian estimates of usefulness prevalence versus polarization of topics. “0” shows that the topic of the video is not polarized while “1” shows the highest polarization.

rank ($C = -3.362$ $p < 0.01$) indicates that the polarization degree of topics has a negative correlation with the prevalence of usefulness. These results also support our findings regarding the time period effect of topics. The usefulness prevalences of some earlier periods (such as 1920s) are lower compared to those whose temporal dimension is later. This is because in these periods the selected topics are more polarized.

We also apply Pearson’s Chi-squared tests between prevalence results of each of the two different topics with various polarization values in our dataset. The results of this study (shown in Table 3.12) indicate that there is statistically significant differences between the prevalence of useful comments between topics with high and low polarization values.

Influence of Entity Types of Topics on Usefulness Prevalence. Our result (shown in Figure 3.5) demonstrates that different platforms (Flickr and YouTube) lead to different usefulness prevalences. For all entity types of topics (place, person, and event), the usefulness prevalence of the Flickr platform is higher than that of the YouTube platform. Furthermore, Figure 3.4 demonstrates that the topic of the media object (event, place, person) leads to different usefulness prevalences. We

get the lowest prevalence of useful comments for topics related to place for both platforms.

For YouTube, topics relating to person have a lower rate of comments than topics related to event. These results concur with our findings in the previous section that the most emotional topic is related to person and the less emotional a comment is, the more useful it is. In contrast, the topics relating to event have the highest rate of useful comments. Events may allow people to give more information about actual places, persons, and happenings. In this way, place and person topics are connected and consequently more information may be given. Contrary to what we expected, the rating results related to the different entity types of topics for Flickr are not similar to the prevalence results for YouTube. For Flickr, the highest prevalence for the three topics, person, place and event, is for person. For topics related to person on Flickr, we recognize that the time periods of many topics of selected photos are earlier compared to the time periods of topics of selected videos related to person for YouTube in our dataset. This is in line with our finding with regard to the effect of time period of topics on usefulness prevalence.

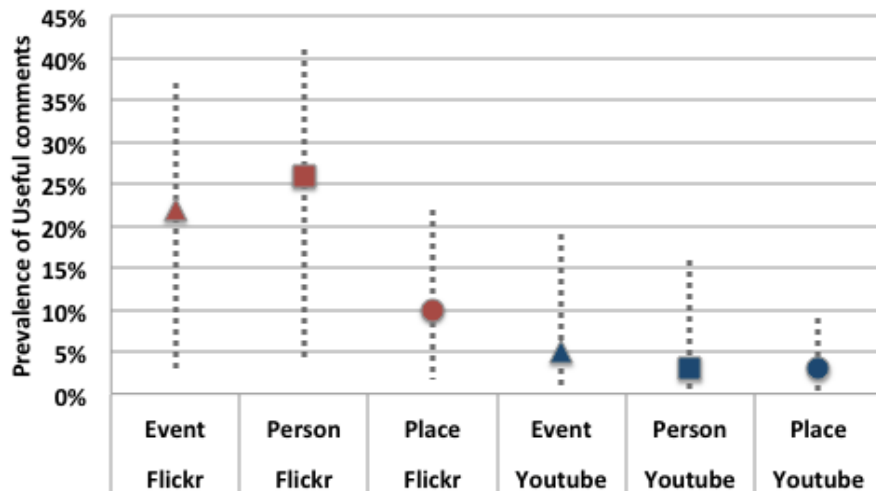


Figure 3.5: Different platforms (Flickr and YouTube) and topics lead to different usefulness prevalence.

Finally, we also apply Pearson's Chi-squared tests between prevalence results of each of the two various topics from different platforms in our dataset. The results of this

Table 3.13: Results of significant differences between the prevalence of useful comments between topics with various polarization values. The numbers indicate X^2 values of Pearson’s Chi-squared tests between each two various polarization values. “*” indicates a significant difference with $p < 0.01$). “**” indicates a significant difference with $p < 0.001$). “***” indicates a significant difference with $p < 0.0001$).

	Flickr-Event	Flickr-Place	Flickr-Person	YouTube-Event	Youtube-Place
Youtube-Person	1801***	191.3***	798.3***	0	98.71***
Youtube-Place	1604***	399.0***	882.2*	98.71***	–
YouTube-Event	1801***	191.3***	798.3***	–	–
Flickr-Person	358.1***	57.52***	–	–	–
Flickr-Place	364.3***	–	–	–	–

study (shown in Table 3.13 indicate that there is statistically significant differences between the prevalence of useful comments between various topics from different platforms.

3.5 Discussion

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

Our experimental findings show that Semantic and Topical features play important roles for inferring the usefulness of comments. For characterizing and inferring the usefulness of comments, a few relatively straightforward features can also be used. Comments are more likely to be inferred as useful when they contain a higher number of references, a higher number of Name Entities, a lower self-reference and affective process (lower sentiment polarity, lower subjectivity tone, swear score, etc). Therefore, we suggest that a design of a platform should urge users to define references [Haslhofer et al., 2010], adding unambiguous users-verified concept references to social media comments. This in turn has a positive impact on the usefulness of comments.

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

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

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

With regard to the analysis of the prevalence of useful comments, our findings indicate that prevalence is influenced by the commenting culture of platforms as well as the different dimensions of topics of media objects. The time period of topics has slight influence on the usefulness prevalence. The nearer the time period of a topic is to the present time, the lower is the prevalence of useful comments. Moreover, the polarization of topics has a negative contribution to the prevalence

of usefulness. This means that for highly polarized topics the prevalence of useful comments decreases. Finally, we find that different platforms (Flickr and YouTube) lead to different prevalences of useful comments. For all entity types of topics (place, person, and event), the prevalence of useful comments on Flickr is higher than that of YouTube, which contains many more non-useful comments.

Chapter 4

Requirements and Design

4.1 Introduction

In the previous sections, we have first discussed the current state of the art in assessment and ranking approaches for user-generated content on the Web, ranging from community-based to single-user assessment and ranking of user-generated content (UGC). Second, we have given an overview of different experiments carried out for the identification of the characteristics of useful comments and creation of a usefulness model. Based on these two steps, we now discuss our novel adaptive moderation framework by describing a number of design considerations and requirements. We introduce the basic concepts that we include in the framework and then give a formal specification of the framework elements by explaining the system architecture and a programming interface specification of the proposed framework. The discussions of this chapter were partially published as a journal article [Momeni et al., 2014b] and are under review for a publication [Momeni et al., 2014a].

4.2 Design Considerations

4.2.1 Fundamental Design Aspects

Considering the results of our experiments and analyzing the state-of-the art, available approaches related to ranking and assessing UGC present a number of fundamental problems, which we need to take into the consideration:

1. *Biases of judgements by crowd*: The wisdom-of-the-crowd approach simply allows all users to vote on (thumbs up or down, stars, etc.) or rate others' content. However, this approach avoids an explicit definition of usefulness. Crowd-based voting is influenced by a number of biases such as “imbalance voting”, “winner circle” (e.g., a “rich get richer” phenomenon), “early bird”, etc. that may distort accuracy [Liu et al., 2007].
2. *Removal of control from end-users*: Many machine-based approaches, which are trained as classifiers to rank comments, are based on a set of majority-agreed labeled comments [Momeni et al., 2013a, Siersdorfer et al., 2010]. 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 requesters to adapt the moderation based on their preferences.
3. *Complexity of usefulness*: Automatic ranking of comments by “usefulness” is generally complex, mainly due to the subjective nature of “useful”. In addition, even human raters find it difficult to agree on the usefulness of comments [Momeni et al., 2013a]. Moreover, usefulness for an individual confounds and blends together two aspects, “relevancy” to what the user has in mind or information she is looking for and “personal interest” in what she attracts her attention. These should be treated separately. For example, a user who intends to look for emotional content may look for comments where the content is relevant to affectivity. However, this does not necessarily mean that this user has any personal interest in the actual comments which are relevant to affectivity. As a result, it is important that systems take into consideration

both these dimensions of usefulness and help individuals adapt ranking based on the particular objective which the user happens to have in mind.

4. *Web UGC as short texts:* UGC are often tiny (such as a tweet, a comment) and they are as fast for users to preview as to read completely. They have no intermediate representation like a headline that can be used for searching and news topic browsing interfaces. Many available approaches propose strategies for extracting topics by enriching the semantics of an individual post [Abel et al., 2011] and enabling users to explore topics in order to filter content with regard to their interest. Topic modeling based methods (both on users and content) feature prominently in this space [Ramage et al., 2010, Chen et al., 2010, Sriram et al., 2010]. Bernstein et al. [Bernstein et al., 2010] 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. FeedWinnower [Hong et al., 2010] is another interface that allows users to rank tweets by various customized factors, such as time and topic. Tseng et al. present a (graph) visualization system called SocFeedViewer [Tseng et al., 2012] that permits users to analyze a topological view of their social feeds.

These approaches address both content and context by learning user preferences and hiding irrelevant content. However, comments are often very brief and topics discussed alongside comments are very noisy. Furthermore, as comments have multiple explicit dimensions (such as language tone, physiological aspects, etc), grouping them exclusively based on topic results in a single imperfect faceted ranking does not enable users to rank comments with regard to other potentially useful facets. Therefore a system which combines higher level features alongside topic classification is desirable.

5. *Various annotating culture in different platforms:* Our results related to usefulness identification experiments demonstrate that different platforms (Flickr and YouTube) lead to different usefulness classification results and the influence of features may vary according to the commenting cultures of platforms. Furthermore, with regard to analysis of the prevalence of useful comments, our findings indicate that prevalence is influenced by the commenting culture

of platforms as well as the different dimensions of topics of media objects. The time period of topics has slight influence on the usefulness prevalence.

Therefore, for a more accurate classification of useful comments, a classification model should be trained with regard to the annotating culture of a platform and media objects.

4.2.2 Conclusions for Design Decisions

In the following, we outline the different aspects that we have considered during the development of our proposed moderation framework. The issues discussed have led us to conclude that the following items are required for the development of the adaptive automated moderation framework:

- A number of strategies for extracting novel facets and topics from UGC that operationalize the complex dimensions of usefulness. These strategies also define the benefits of combining different types of facets (such as facet related to topic, subjectivity, etc) for providing end-users with access to interesting or relevant comments.
- An interactive framework for leveraging these facets to directly enable end-users to adaptively moderate UGC based on their preferences and interests with regard to the commenting culture of the platform.
- A possibility for users to provide feedback simultaneously by implicit means (using the faceted browser) or explicit means (voting). Both of these can be utilized to build user models and improve the automated moderation processes.
- A possibility of assessment of usefulness without users' feedback. While it is preferred that the feedback is provided by the user, it is helpful to begin with a "baseline" assessment of usefulness that is independent of the user.

With regard to the issues and requirements discussed, this thesis proposes an alternative, *automated* support for the *multi-faceted adaptive ranking* of user-generated

content. The proposed framework, which is influenced by past work on multifaceted search [Koren et al., 2008], active learning, and topic identification is a semi-supervised learning approach for adaptive ranking of social media UGC with regard to the preferences of each individual user. The proposed ranking framework clusters each element of a comment along multiple explicit semantic facets (e.g., subjective comments, informative comments, and topics, etc). This enables the clustering to be accessed and ordered in multiple ways rather than in a single, topic order [Bernstein et al., 2010, Abel et al., 2011], and also avoids having to rely on particular majority-agreement sources of ground truth. Although, the core component of the framework is a baseline usefulness prediction model which is trained based on majority agreement of users for useful comments. The system uses this model as the baseline if the user does not explicitly or implicitly give the system feedback. For adaptive moderation, starting from a possibly empty set of manually labeled comments, a machine-based algorithm semantically enriches comments, provides clusters of comments, and accordingly proposes relevant facets. Users explore different clusters (different facets such as topics discussed among comments, subjective opinion, etc) and select combinations of facets in order to rank and extract comments that match their interests or are relevant to selected facets.

Furthermore, the framework allows end-users to interactively (e.g., by providing an augmented Web-based user interface) rank comment feeds based on their interests both through implicit and explicit feedback. Implicit feedback is related to facet selection behavior of the user. Explicitly given feedback by a user (labeling comments) are used by the system for training and improving clustering models and user models. For capturing both dimensions of usefulness, personal interest and relevancy, users are given the chance to provide two explicit labels (votes): “Relevant” and “Interesting”. “Relevant” votes capture how the comment is related to an information need by the user, whereas an “Interesting” vote capture the user’s personal interest.

Given the basic description of the proposed approach and the challenges presented in the context of social media object sharing platforms (such as YouTube, Flickr), we provide a detailed description of the system architecture of the proposed framework in the following.

4.3 AMOWA: A Framework for Adaptive Moderation of UGC

This work proposes a framework, AMOWA (Adaptive Moderation of Web Annotations), which provides automated support for adaptive faceted ranking of user-generated content on social media, thereby helping users to explore content and personally identify useful content in accordance with their preferences. High level operational processes of the framework shown in Figure 4.1. To enable this, the framework enriches content provided by a user along different semantic facets (e.g., subjectivity, emotional level, and topics) and actively learns from both implicit and explicit actions by users. Implicit actions include the browsing behavior of the end-users as they make facet and topic selections, explicit actions include directed voting by end-users on comments. The proposed framework comprises four main components:

4.3.1 Component 1: Semantic Enrichment

Comments are often short and do not explicitly feature facets which describe their content. Our proposed framework first enriches each comment along various semantic facets. This component utilizes two core strategies to enrich comments: (1) topic-based enrichment using extracted named entities (NEs) and (2) feature-based enrichment where comments are automatically characterized by a set of semantic facets. We categorize proposed facets by framework into three broad types of facets: Topic-related facets **TF**, Subjective facets **SF** (such as comments with subjective tone, highly affective language, offensive and anger oriented, sad oriented, religion referenced, etc. by utilizing Subjectivity Lexicon [Wilson et al., 2005] and LIWC¹ [Tausczik and Pennebaker, 2010]), and Objective facets **OF** (such as informative, video timestamp, etc.). Table 4.1 shows an overview of our proposed facets. Related works and observations on available approaches for analyzing free-text user-generated content in online communities and social media encourage the use of these facets. For example, the dichotomy of objectivity or subjectivity of comments is mo-

¹Linguistic Inquiry and Word Count Lexicon

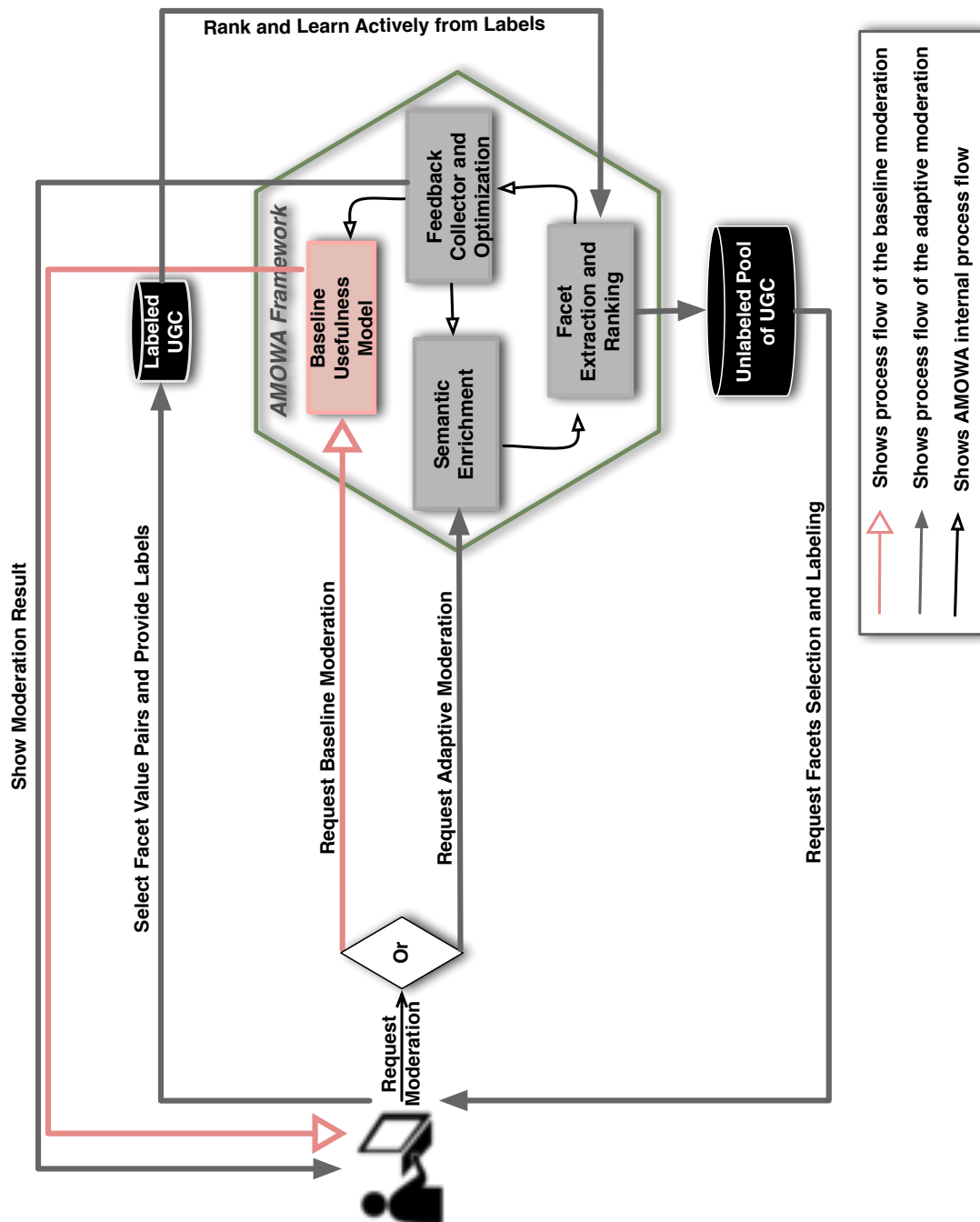


Figure 4.1: Abstract overview of proposed adaptive moderation framework.

tivated by a work proposed by Diakopoulos et al. [Diakopoulos and Naaman, 2011]. It is important to note that the set of facets which we explore in this work is a minimum set of facets to demonstrate the effectiveness of using adaptive faceted ranking but is not necessarily a complete set of useful facets.

Formally, let's define a set of N comments $C_m^{ordinary} = \{c_i\}_{i=1}^N$ on a media object m . A semantic enrichment function, E , enriches each comment $E : C_m^{ordinary} \rightarrow C_m^{enriched}$. Similarly, $C_m^{enriched} = \{(c_i, x_i)\}_{i=1}^N$ is a set of N enriched comments, where for each comment c_i , $x_i \in \mathbb{R}^{|v|}$ is the comment's semantic facet vector representation, with facet space of size $|v|$.

Identifying the topic of a short text, such as a comment, is difficult. By experimenting with a number of approaches (cf. study 3, Section 6.4), we find that the Named Entities-based approach is a useful proxy for identifying the topics of the comments and thus it is used by this component.

4.3.2 Component 2: Facet Extraction and Ranking

This component operates on semantically enriched comments and clusters comments along multiple explicit semantic facets and then selects a list of facets dynamically. Furthermore, this component enables an individual user to explore facets, select a combination of facets, and rank comments accordingly.

For clustering purposes, we utilize the centroid clustering method on enriched comments (those annotated by the Semantic Enrichment component) by receiving a number of all facets that belong to a media object. For a dynamic selection of facets, we choose **Greedy Count** as a simple algorithm, which is also considered in [Lieberman and Lempel, 2012]. It ranks the facets to be selected according to the number of top- X comments in the ranking result set and is similar to the Most Frequent heuristic selection used by [Koren et al., 2008]. The Greedy Count approach favors more popular facets and is likely to result in many drill downs, as the total number of comments that will be filtered with each click will be relatively small.

Let F denote the set of all facets and $F_m \subseteq F$ represents all facets that belong to $C_m^{enriched}$. The main use case that we consider is that a user submits a media object

Facet	Type	Description
<i>Subjectivity Tone</i>	SF	Measures the subjectivity degree of a comment. The Subjectivity Lexicon [Wilson et al., 2005] is used to calculate subjectivity.
<i>Offensive and Angry</i>	SF	The extent to which an author uses offending language or the comment reflects the author’s anger. Approach relies on the LIWC [Tausczik and Pennebaker, 2010].
<i>Sad</i>	SF	The extent to which the comment reflects the author’s sadness. Approach relies on the LIWC [Tausczik and Pennebaker, 2010].
<i>Self-reference</i>	SF	The extent to which an author refers to herself, e.g. “I”, “mine”. Approach relies on the LIWC [Tausczik and Pennebaker, 2010].
<i>Affective</i>	SF	Positive and negative sentiments. $SenPolarity = \frac{\#Positive\ Words + \#Negative\ Words}{\#Words}$
<i>Informativeness</i>	OF	Novelty of terms of a comment compared to other comments on the same object.
<i>Text Statistics</i>	OF	Number of words, verbs, adverbs, etc.
<i>Linkage Variety</i>	OF	Number of hyperlinks in a comment.
<i>Religious Referenced</i>	OF	The extent to which an author employs religion-oriented word such as “mosque”, “church”, etc. Approach relies on the LIWC [Tausczik and Pennebaker, 2010].
<i>Video Timestamp</i>	OF	The extent to which a comment points to a part of a video. Extracting Entities related to time.
<i>NE Types Variety</i>	TF	Number of distinct types of named entities.
<i>Named Entities</i>	TF	Number of named entities.

Table 4.1: Overview of our proposed facets. For each facet, we also provide the facet type: Subjective (**SF**), Topic-related (**TF**), and, Objective (**OF**). Each facet is extracted from each comment.

m to the ranking framework. Next, the framework computes a list of enriched comments, $C_m^{enriched}$ and selects and displays a set of facets $F_p \subseteq F_m$ of size l . Thus, S represents the facet selection function, where $S : C_m^{enriched} \rightarrow F_p$ maps a set of enriched comments to a subset of facets.

Finally, a user is enabled to explore facets, select a combination of facets, and rank comments accordingly. For ranking comments, multiple selections of facets are treated as an ‘and’ rather than an ‘or’. This means that the facets’ values are combined conjunctively for ranking comments.

4.3.3 Component 3: Feedback Collector and Optimization

The goal of this component is to enable users to provide implicit and explicit feedback. This feedback facilitates evaluation of different strategies related to various facet types, and, furthermore, it facilitates the optimization of facet selection.

Implicit activities of users in the system such as facets’ exploration and selection 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 “Interesting”. We use 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 the end-user’s perspective (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).

More formally, let F_s denote a set of selected facets in a previous ranking by a user, and $C_{m,labeled}$ is a set of comments labeled by the user. This set contains two subsets: $C_m^{relevant} = \{(c_i, r_i)\}_{i=1}^N$ is a set of comments labeled as relevant by the user, where for each comment, $r_i \in \{0, 1\}$ gives the comment’s label (0 for irrelevant, 1 for relevant) and $C_m^{interesting} = \{(c_i, t_i)\}_{i=1}^N$ is a set of comments labeled as interesting by the user, where for each comment, $t_i \in \{0, 1\}$ gives the comment’s label (0 for non-interesting, 1 for interesting).

We now express the facet selection optimization approach as the reduction in ranking effort in terms of the proposed facets $F_{p,s}$. For this purpose we adopt the approach

taken in [Lieberman and Lempel, 2012] and modify it. We can define the utility of displaying a set of facets F_p proposed by M , a facet’s optimization approach, with respect to m , a media object, F_s , a set of already selected facets, and $C_{m,labeled}$, a set of labeled comments, as follows:

$$U_{m,F_s}^M = E[X|m, F_s] - E_M[X|m, F_s, F_p]$$

$$E[X|m, F_s] = \sum_{\substack{c \in C_{m,labeled} \\ r_c^{F_s}(c) > m}} p(c = c_i | t_i = 1) r_c^{F_s}(c)$$

where $E[X|m, F_s]$ represents the expected ranking effort of a user that does not click on facets for the media object, $E_M[X|m, F_s, F_p]$ represents the ranking effort using the facets proposed by approach M , and X denotes the random variable that represents the search effort of a user for one click. $r_c^{F_s}(c)$ denotes the rank of c in the result set, and $p(c = c_i | t_i = 1)$ is the probability of c being an interesting comment.

Using this definition, we can formulate facet selection optimization approach as:

$$F_{p,s,M}^* = \arg \max_{\substack{F_p \subseteq F \\ |F_p| < k}} U_{m,F_s}^M$$

where k is the size of the facet subset to be shown to users. This optimization approach is NP-hard and therefore it is challenging to have an exact optimal solution to this problem – reduction from the Hitting Set problem [Lieberman and Lempel, 2012]). In this work, we explore the optimization of facet selection in terms of investigating the strategies for selecting the various types of facets in Section 6.2.

In addition, this component can use this feedback for active learning which permits: (1) creating a user model and personalized ordering and selection of facets and accordingly extraction of comments in accordance with user’s interests, and (2) improving the clustering and ranking of comments. For example, the framework can use “Interesting” votes to create a user model and can also use “Relevant” votes to assess and improve the performance of the Facet Extraction and Ranking component. The user model can be used by the system for creating a user interest profile that represents the current interest and actions of the user so that the faceted ranking component can use this profile in order to optimize personalized selection of facets. (for example, if a topic in the user’s interest profile appears with a higher

rate, a topic of great interest will be promoted to the top of the facet set even if there is only a single comment). It is important to note that there are many ways to construct and leverage user models for personalized ranking. Depending on the explicit (and implicit feedback) structure, these can be quite varied.

However, this work primarily focuses on examining different strategies for semantic extraction of facets and adaptive ranking. We primarily use the feedback provided by this component to evaluate the performance of different facet selection strategies and do not explore the strategies for creating and using user models.

4.3.4 Component 4: Baseline Usefulness Model

While we prefer the feedback provided by the user, it is helpful for us to begin with a “baseline” moderation that is user-independent. The system uses the baseline “usefulness” classifier if the user does not explicitly or implicitly give the system feedback. This model predicts whether each unlabeled comment is useful or non-useful. This means using the labeled training comments, $C^{labeled}$, learn a supervised “Usefulness” classifier, which identifies the usefulness of each comment in $C^{enriched}$, $f : \mathbb{R}^{|v|} \rightarrow \{0, 1\}$.

Following our experiment for developing a usefulness prediction model (see Section 3.4.1), a small amount of majority-agreement on labeled training comments may be assumed for training a usefulness classifier. The Logistic Regression model trained on a particular set of semantic features (see Table 3.9) performs well when identifying comments that may be considered useful by a majority of users. When trained on useful and non-useful comments, this classifier is found to perform well relative to a baseline (predicts usefulness using only the Subjectivity Tone, which is a particularly strong baseline as a result of the feature analysis study for usefulness by Momeni et al. [Momeni et al., 2013a]). This component covers the baseline moderation process shown in Figure 4.1.

4.4 Functional Specification of AMOWA by Means of Services

In this section, we describe various interfaces that translates the AMOWA framework into an implementation-centric specification and define functional specification of the proposed framework. This can be taken as a reference for implementations. These interfaces are designed by means of different services. Then in chapter 5, we show how the presented specification can be used in the context of an application scenario such as a Web service with an interactive user interface.

We divide the specification of the AMOWA interface specification into three parts. In the first part, “*SEFE (Semantic Enrichment and Facet Extraction) Service*”, we describe how the framework semantically enriches content and selects a set of facets, which is related to two components: “Component2: Semantic Enrichment” and “Component3: Facet Extraction and Ranking” . In the second part, “*Ranking Service*”, we describe how the framework enables the user to select facet-value pairs (FVPs) and then ranks content with regard to selected FVPs or with regard to a core usefulness model without selecting FVPs, which related to “Component1: Baseline Usefulness Model” and ranking function of the “Component3: Facet Extraction and Ranking”. In the third part “*Feedback Service*” we describe how the framework takes a list of comment label pairs (CLPs) and creates or updates a core moderation model related to a user, which is related to “Component 4: Feedback Collector and Optimization”.

4.4.1 Interface Specification — SEFE Service

“SEFEService” is an interface that takes as an input a list of comments on a media object, semantically enriches comments, and selects relevant facets. When “SEFEService” calls “InputLoader” and “FacetComposite”, it crawls all information related to comments and commentators of a media object and creates a “SEFE” data object. “SEFE” is a data object for accessing a “Corpus” data object and a list of sorted “Facet” data objects for all comments on the input media object

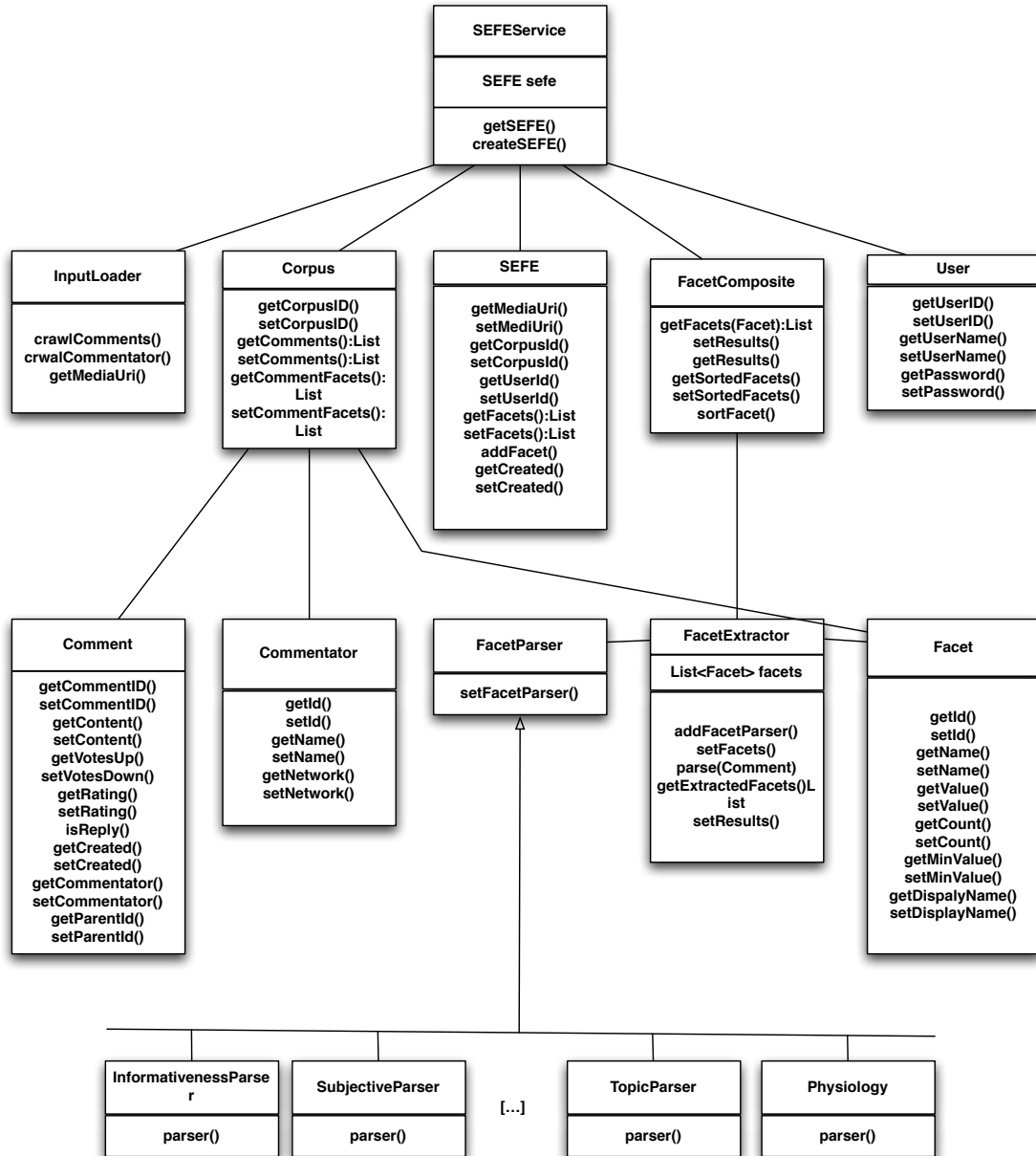


Figure 4.2: Interface specification of SEFE service

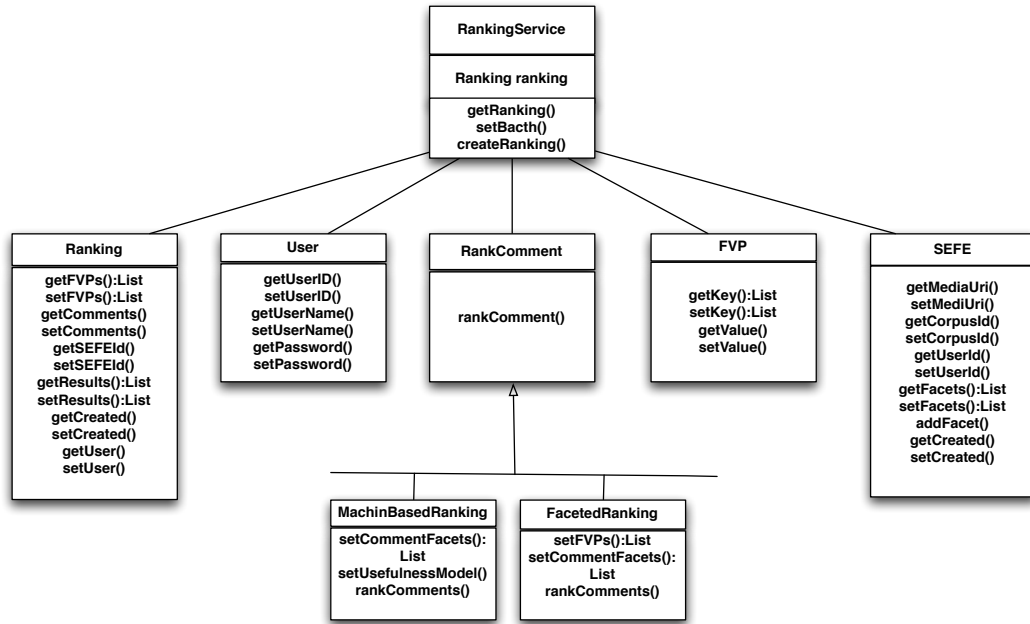


Figure 4.3: Interface specification of Ranking service

and “User” data object. “User” is a data object for accessing information related to users who call “SEFEService”. “Corpus” is a data object for accessing a list of “Comment” data objects and connections between comments and their related facets. “Comment” is a data object for presenting the content of a comment, a reply tree of the comment, and other relevant information to the comment. “Input-Loader” is an interface for crawling information related to comments and commentators and for storing characteristics of a comment as a “Comment” data object. “FacetComposite” is an interface which uses “FacetedExtractor” for extracting a list of “Facet” data object from a list of comments. “Facet” is a data object which stores characteristics of a facet. “FacetedExtractor” is an interface which calls different “FacetParser” interfaces, such as “InformativenessParser” for enriching a comment with a high informativeness score or “TopicParser” for enriching and extracting topics from a comment. Figure 4.2 shows interface specification of the SEFE service.

4.4.2 Interface Specification — Ranking Service

“RankingService” is an interface that takes as inputs a list of facet value pairs (FVPs) and a set of enriched comments, and ranks comments with regard to selected FVPs. “RankComment” is an interface based on a list of “FVP” data objects. It ranks enriched comments — provided by “SEFE” data object — and creates a “Ranking” data object. “FVP” is a data object which stores information related to a facet value pair. “Ranking” is a data object which stores information related to ranking results. The “RankComment” interface inherits two other interfaces: “MachinBasedRanking” and “Facetedranking”. If user does not select FVPs, then the framework uses “MachinBasedRanking” which uses the baseline usefulness prediction model (see Section 4.3.4) to rank comments. Instead, if a user select FVPs, then the framework uses “Facetedranking” for ranking comments. Figure 4.3 shows interface specification of the Ranking service.

4.4.3 Interface Specification — Feedback Service

“FeedbackService” is an interface that takes a list of comment label pairs (CLPs) and creates or updates a core moderation model related to a user. When “FeedbackService” calls “UserModelCreator”, it collects a list of “CLP” data objects. Based on the collected label for a “Corpus” data object, it creates or updates a “UserModel” data object. “UserModelCreator” is another interface and when it calls “LabelComment”, it collects labels for set of comments (a Corpus object). “LabelComment” is an interface that adds a label for a comment by a user and creates a “CLP” data object. “CLP” is a data object that stores and presents a related label for a comment given by a user. “UserModel” is a data object which stores a list of “CLP” objects given by a user and a baseline moderation model. This baseline moderation model is a model based on machine learning algorithms and is trained based on majority-based judgments (see 4.1). In addition, it is actively trained and updated by labels gathered from a user. Figure 4.4 shows interface specification of the Feedback service.

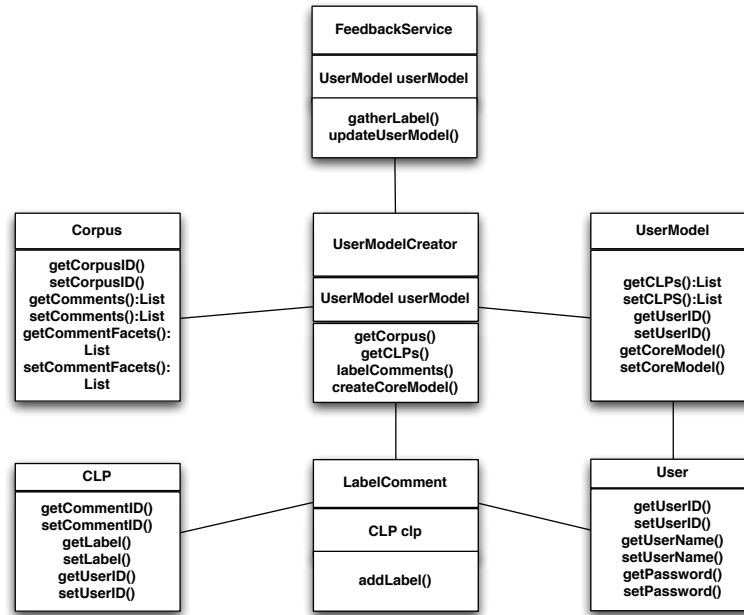


Figure 4.4: Interface specification of Feedback service

4.5 Summary

The results of our experiments and a state-of-the art analysis indicate that available approaches related to ranking and assessing UGC exhibit a number of fundamental problems: biases of judgements by crowd, removal of control from end-users, complexity of usefulness, Web UGC as short texts, and various annotating cultures in different platforms. These issues lead us to conclude that the following aspects are required for the development of the automated moderation framework: a number of strategies for extracting novel facets and topics from UGC that operationalize the complex dimensions of usefulness, an interactive and adaptive framework for leveraging these facets to directly enable end-users to moderate UGC based on their preferences and interests, a possibility for users to provide feedback simultaneously by implicit or explicit means, and finally a possibility of assessment of usefulness without users' feedback.

With regard to design considerations and requirements discussed above, we have

introduced the AMOWA framework and a interface specification of AMOWA by means of services. The static part of the AMOWA interface specification consists of a set of types that reflect all components of the AMOWA model. The interface specification can be easily transformed to any object-oriented language because it has been specified in a generic UML notation.

Chapter 5

Implementation

5.1 Introduction

After having presented the AMOWA framework and its elements in various levels of abstraction, we now discuss prototypical implementations of the most important parts of the proposed framework. By focusing on these parts, we can show the flexibility of the AMOWA framework and its applicability to different social media platforms. Three implementations are examined:

1. *Baseline Usefulness Model*: we discuss a prototypical implementation of the baseline model for automatically predicting usefulness of UGC without receiving explicit or implicit users' feedback. This prototype covers the baseline moderation process shown in Figure 4.1.
2. *AMOWA-WS*: we discuss a prototypical implementation of a Web service of AMOWA which can be simply integrated as a plugin into any social media platform or any platform which deals with UGC. It enables end-users to moderate content with regard to their personal interest or task in hand.
3. *AMOWA-UI*: we discuss an implementation of a Web user interface of AMOWA. This user interface is a client-site implementation of the AMOWA-WS which

allows users to access AMOWA-WS and work with the moderation framework using interaction metaphors.

In the following, we outline the architecture and important implementation aspects of each of these prototypes.

5.2 Usefulness Prediction Model

While the framework prefers the feedback provided by the user, it is helpful for us to begin with a “baseline” assessment of usefulness that is user-independent. As discussed in Section 4.3.4, the baseline component of the framework is the usefulness classifier which predicts whether each unlabeled comment is useful or non-useful without using feedback from users. Following our usefulness experiments described in Section 3.4, a small amount of majority-agreement on labeled training comments may be assumed. Therefore, by using Weka¹— a machine learning software written in Java — a usefulness classifier is trained using a supervised learning algorithm on the manually coded comments from two datasets crawled from two social media platforms (1000 useful comments and 1000 not useful comments from the Flickr data, and 400 of each class from YouTube). The Logistic Regression model trained on a particular set of semantic features (see Section 3.2) performs well when identifying comments that may be considered useful by a majority of users.

For developing semantic features related to subjectivity, topic, and psychological characteristics of the text of a comment, three semantic “*enrichment components*” are developed using the Java programming language (version Java 1.6): (1) for psychological content characteristics, we develop an API using Linguistic Inquiry and Word Count lexicon [Tausczik and Pennebaker, 2010], (2) for subjectivity, we develop an API using Subjectivity Lexicon [Wilson et al., 2005], and (3) for Named Entities related features, we develop an API using Gate toolkit². These components are also used for extracting facets from a corpus of comments.

¹<http://www.cs.waikato.ac.nz/ml/weka/>

²<https://gate.ac.uk/projects.html>

When this classifier is trained on comments labeled useful and non-useful from YouTube and Flickr, it is found to perform well relative to a baseline which predicts usefulness using only Subjectivity Tone. This baseline is particularly strong as a result of the feature analysis study for usefulness, presented in Chapter 3. The classifier also performs for Flickr with a precision of .87 and recall of .9 (compared to .65 and .8 for the strong baseline) and YouTube with a precision of .65 and a recall of .83 (compared to .55 and .7 for the strong baseline). As performance of the classifier is lower for YouTube, the YouTube dataset is evaluated using our proposed adaptive moderation framework, as described in Section 6.

5.3 AMOWA–WS (A Web Service for AMOWA)

This section discusses prototypical implementations of a Web service with regard to the proposed interface specification of the framework which is presented in Section 4.4. This Web-based interface enables end-users to moderate social media content based on their preferences and interests when using the Web.

The Web service is implemented using Java programming language, MongoDB (as backend dataset), and JAX-RX. The structure of the project changed frequently in its initial stages due to the different facets and features which were implemented. Therefore, we used MongoDB because of its persistency. JAX-RS is a specification of an accumulation of Java-APIs for implementing REST style Web service with Java. One implementation of this specification is Jersey which was used in version 1.179. This version is based on the JAX-RS specification 1.1.

In order to enable the integration of the Web service with any type of social media platform, the Web service is developed to receive a set of comments on an online media object (such as news article, YouTube Video, etc) in XML format with minimum and simple schema. Therefore, any platform can convert its content into the required format and requests for the moderation process.

With regard to the interface specification of the AMOWA, which is divided into three parts: “*SEFE Service*” (how the framework semantically enriches content and extracts facets), “*Ranking Service*” (how the framework enables the user to select

Table 5.1: Description and arguments of the Web-based interface related to SEFE service

POST /sefes		
Description		Receives a corpus of comments and creates a SEFE object which contains a list of enriched comments and a list of related facets. This means it enriches comments and extracts a list of related facts.
Arguments	file	Input file which contains a list of comments and information regarding comments such as the author of a comment and the media object. This input file is currently represented in XML data format.
Returns		SEFE object (see Section 4.4.1)
Field Description:	corpusId	The ID of the corpus of comments.
	created	Creation date of the SEFE object.
	facets	Extracted facets from the corpus.
	facets.count	Quantity of the facet inside the corpus.
	facets.facetId	ID of the facet.
	facets.minValue	Minimum value of facet to become counted in the corpus (server side setting).
	facets.mediaUrl	URL of the media object.

facet-value pairs (FVPs) and then ranks content with regard to selected FVPs or with regard to a core usefulness model without selecting FVPs.), and “*Feedback Service*” (how the framework takes a list of comment label pairs (CLPs) and creates or updates a core moderation model related to a user), accordingly, the specification of the Web service follows the same division.

Table 5.1 shows description, arguments, and output fields of the Web-based interface related to SEFE service. By uploading a corpus of comments, using an input file, which is currently represented in XML data format (List 5.1 shows structure and required fields for an input XML file), this service semantically enriches comments using three “Semantic Enrichments”, APIs (see Section 5.2), extracts a list of se-

lected comments based on frequency, and provides a SEFE object which is currently represented in the JSON data format (List 5.2 shows structure and output fields for the output representation). In the output representation, fields which are named “count” show the percentage of comments related to the respective facet or topic. These numbers are then used to rank and adapt the ordering of facts for AMOWA–UI. Using an input file — which can be represented in any data format — enables us to integrate the service with any platform related to user-generated content, such as YouTube, Flickr, Twitter, etc. This means the service can be connected to the API or the data feed of any platform and transform content into the required data format. Consequently, the service can be easily integrated in any other platform.

Listing 5.1: A Sample of an input XML file format

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <MediaComments mediaUrl="http://www.youtube.com/watch?v=e65XLPSDXD4">
3   <Comment id="38" author_id="1" created="1294862756119">
4     <content> Mostly because President
5       Franklin D. Roosevelt realized that if
6       prohibition was repealed, farmers could
7       sell more grain; carpenters could build
8       more kegs; breweries, distilleries and
9       bars would re-open and employ more
10      people; and the US Gov't could tax
11      the liquors being produced and consumed
12      (The USA was in a Depression then).
13    </content>
14  </Comment>
15  <Comment id="39" created="1294863656114" parent_id="38">
16    <content>Wrong...44?
17  </content>
18  </Comment>
19  <Comment id="40" created="">
20    <content>Ok.? I'm 51 now. Looking back on
21      it now. I was 13 at the time.
22      Nixon resigned over THIS stupid
23      crapola? Help! I'm losing my
24      country.
25    </content>
26  </Comment>
27 </MediaComments>
```

Table 5.2: Description and arguments of the Web-based interface related to the Ranking service

GET /rankings		
Description		Ranks a corpus of comments based on selected facet value pairs.
Arguments	sefeId	ID of SEFE object (string). A list of enriched comments and related extracted facets.
	fvp	Facet Value Pairs (e.g: fvp=swear:0.0,affect:0.0,topic:nixon) (array).
	limit	Optional argument for limiting a maximum number of ranked comments within this ranking (integer).
Returns		Ranking object (see Section 4.4.2)
Field Description:	sefeId	The ID of the input SEFE object.
	created	Creation time of the ranking.
	comments	Comments which match the Facet Value Pair selection.
	comments.\$.content	The texts of comments.
	comments.facets.\$.facetId	ID of the facet.
	comments.\$.hasJudgement	True or false if the comment has a judgement.
	comments.\$.id	ID of the comment (as defined in the input file).
	comments.\$.myParent	The ID of the parent of the comment.
	fvp	Selected Facet Value Pairs for this ranking.
	results	Results of ranked comments within this ranking.

Table 5.3: Description and arguments of the Web-based interface related to the Feedback service

GET /judge		
Description		Sets a judgment field and the desired values for ranked comments.
Arguments	rankingId	ID of Ranking object (string).
	commentId	ID of comment which will be judged (labeled).
	field	The type of judgments which are Interesting or Relevant (string).
	value	True or false (string).
Returns		Modified Ranking object which includes judgments for comments.

The next part of the AMOWA-WS deals with the “Ranking Service”. Table 5.2 shows description, arguments, and output fields of the Web-based interface related to the Ranking service. By receiving a list of comments — setting safe ID — and a set of facet value pairs, this interface ranks comments with regard to selected FVPs and provides a ranking object which is currently represented in JSON data format. List 5.3 shows an example of such JSON output.

Finally, the last part of the AMOWA-WS deals with the “Feedback Service”. Table 5.3 shows description, arguments, and output fields of the Web-based interface related to the Feedback service. By receiving a list of ranked comments, types of judgments which are “Interesting” or “Relevant”, and values of judgement which are “False” or “True”, this interface adds a label for a comment and provides a modified ranking object which also includes labels for comments.

All requests except login and register may include BasicAuth Header parameters for (simple) user identification. This is due to the fact that the framework requires storing and tracking users’s explicit and implicit feedback. Table 5.4 describes an interface for creating a “User” object and shows arguments and fields required for creating user ID. List 5.4 shows an output of this service using JSON data format. The password is encoded with Base64 ³.

³Authentication is used with the following string: “Basic username:base64encoded — password”

Table 5.4: Description and arguments of Web-based interface related to Login service

Post /login		
Description		Logs in a user
Arguments	username	username (string)
	password	password (string)
Returns		The User object

Listing 5.2: A sample of a JSON file format of SEFE object

```

1 {
2   id: "52ea239b036472844d65bd8f",
3   version: "1",
4   corpusId: "52ea2399036472844d65bd81",
5   created: "2014-01-30T11:04:11.854+01:00",
6   facets: [
7     {
8       count: "90.0",
9       facetId: "topic",
10      minValue: "0.0",
11      value: {
12        {
13          @type: "topics",
14          values: [
15            {
16              count: "23.076923076923077",
17              name: "Nixon"
18            },
19            {
20              count: "7.6923076923076925",
21              name: "president"
22            }
23          ]
24        },
25        {
26          count: "84.61538461538461",
27          facetId: "subjectivityNormal",
28          minValue: "0.0",
29          value: {
30            @type: "xs:double",
31          }
32        },
33      ]

```

```

34         count: "76.92307692307692",
35         facetId: "affect",
36         minValue: "0.0",
37         value: {
38             @type: "xs:double",
39         }
40     },
41 ],
42     mediaUrl: "http://www.youtube.com/watch?v=e65XLPSDXD4",
43     userId: "notlogged"
44 }

```

Listing 5.3: A sample of Ranking object in JSON data format

```

1  {
2      id: "52f89eaa0364fb4b1a98e725",
3      version: "1",
4      sefeId: "52f89c910364fb4b1a98e71e",
5      comments: [
6          {
7              content: "Hong Kong has nothing to do with whether you lot
8              imperialists or not. 1997 was the year in the unequal
9              treaty of Nanking that Hong Kong was to be returned.
10             And to be honest do you think the UK will fight for the
11             colony? against China? Yeah you might be able to win
12             against Argentinians in Falkland, but against China? I
13             don't think so. Basically the UK HAD TO return Hong Kong in
14             1997 and had nothing to do with being imperialistic anymore
15             or not.",
16             facets: [
17                 {
18                     "facetId: "sad",
19                     "value: "0.002079002079002079"},
20                 { "featureId" : "affect",
21                   "value" : 0.006237006237006237 },
22                 { "featureId" : "subjectivityNormal",
23                   "value" : 0.01098901098901099 },
24                 { "featureId" : "anger",
25                   "value" : 0.002079002079002079 },
26                 { "featureId" : "topic",
27                   "values" : [ "Hong kong", "China",
28                             "England", "1997", Argentinian] }
29             ]

```

```

30         ],
31         hasJudgement: "false",
32         id: "39",
33         _Id: "52f89c900364fb4b1a98e712"
34     }
35 ],
36 comments: [
37     {
38         content: "I was born in Hong Kong and relocated in the US
39         when I was 8, many other people immigrated to the US or
40         Canada to get away from China prior to 1997. I have always
41         said that the British Colonization is what made Hong Kong
42         what it is. I hope China would adopt the ways of Hong Kong
43         instead of force the ways of China into Hong Kong. With the
44         current 50 years one country two systems rule, I'm still
45         unsure what their future would be like.",
46         facets: [
47             {
48                 "facetId": "selfReference",
49                 "value": "0.004587155963302753"},
50             { "featureId" : "affect",
51               "value" : 0.01146788990825688 },
52             { "featureId" : "anger",
53               "value" : 0.002293577981651376 },
54             { "featureId" : "sad",
55               "value" : 0.002293577981651376 },
56             { "featureId" : "topic",
57               "values" : [ "Hong kong", "China",
58               "Usa", "Canada", "China",
59               "1997" ] }
60         ]
61     },
62     hasJudgement: "false",
63     id: "40",
64     myParent: {
65         content: "yes i was 13 too",
66         hasJudgement: "false"
67     },
68     _Id: "52f89c900364fb4b1a98e713"
69 }
70 ],
71 created: "2014-02-10T10:40:58.464+01:00",
72 fvp:

```

```

73 [ { "key" : "informative", "value" : "0" },
74   { "key" : "topic", "value" : "Hong kong" } ]
75 results: "2"
76 }

```

Listing 5.4: A sample of a JSON file format of User object

```

1 {id: "52ea31d5036472844d65bd99"
2   password: "test123"
3   username: "test" }

```

5.4 AMOWA–UI (A User Interface for AMOWA)

This section discusses prototypical implementations of a user interface of the proposed AMOWA-WS with regard to the proposed interface specification of the framework, presented in Section 5.3. This user interface allows users to access the Web service and work with the moderation framework using interaction metaphors.

In the design and implementation of the AMOWA–UI, we focused on three main design objectives that have been outlined by Hearst [Hearst, 2009] as important constructs for the development of user interfaces like AMOWA–UI: effectiveness (helpful to rank UGC with regard to user’s interest and the relevancy to the user’s particular objective and task), efficiency (quick to process information, enabling users to extract useful information quickly), and satisfaction (easy to use and browse, enjoyable, overwhelming, or tedious).

The user interface is written in CoffeeScript and HTML5. It uses Backbone.js⁴ as a library for structuring the front–end application and jQuery for DOM-interactions.

This UI (see Figure 5.1) enables end-users to perform the faceted adaptive ranking of social media comments (such as comments on YouTube videos or online news articles) and allows their performance to be evaluated. In general, the faceted ranking framework user GUI includes two parts, one displaying the facets, and another displaying the ranked results (see Figure 5.1). Based on such an interface, a user can perform different actions:

⁴www.backbonejs.com

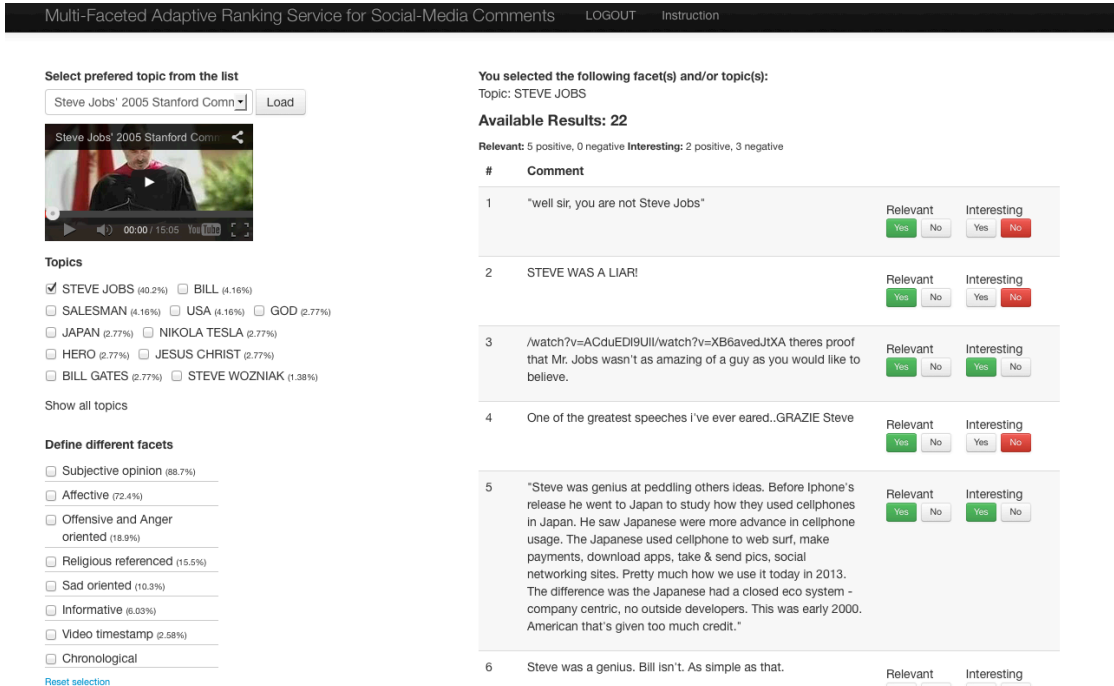


Figure 5.1: Screenshot 1 of the Web-based user interface of the framework

- A user entering a media object ID triggers the system to crawl all comments related to the media object, semantically enrich each comment along multiple semantic features, cluster each comment with regard to the value of its features into coherent facets, and finally show a list of facets and topics on the left side of the user interface.
- A user selecting combinations of proposed facets based on her preferences triggers the system to show ranked lists of comments based on user's selections.
- A user browses comments and votes if the comments matches her interests or are relevant to her selections of facets.

The system provides three types of facets: TF–Topic-related facets, SF–Subjective facets (such as comments with subjective tone, highly affective language, offensive and anger oriented, sad oriented, religion referenced, etc), and OF–Objective facets (such as informative, video timestamp, etc). Some examples of these facets are shown in Table 4.1.

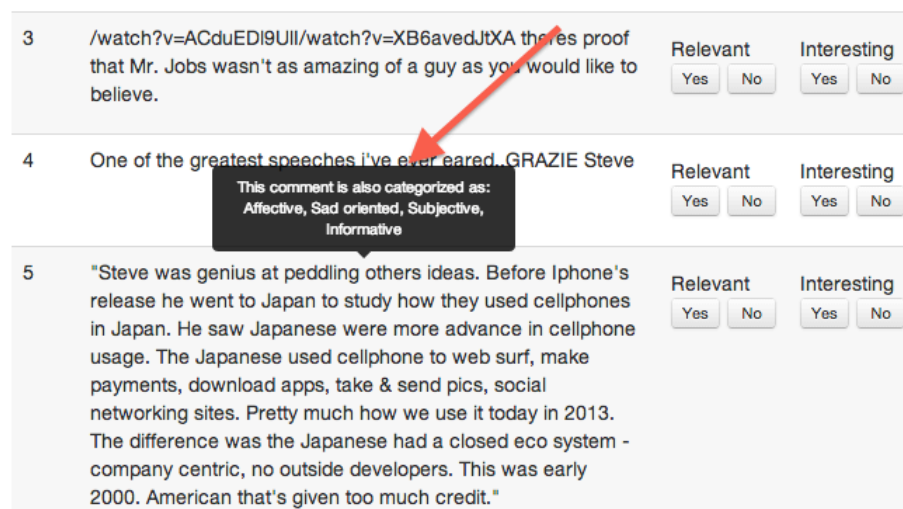


Figure 5.2: Screenshot 2 of the Web-based user interface of the framework. The system also shows a so-called on-fly short overview of all other possible facets for each comment.

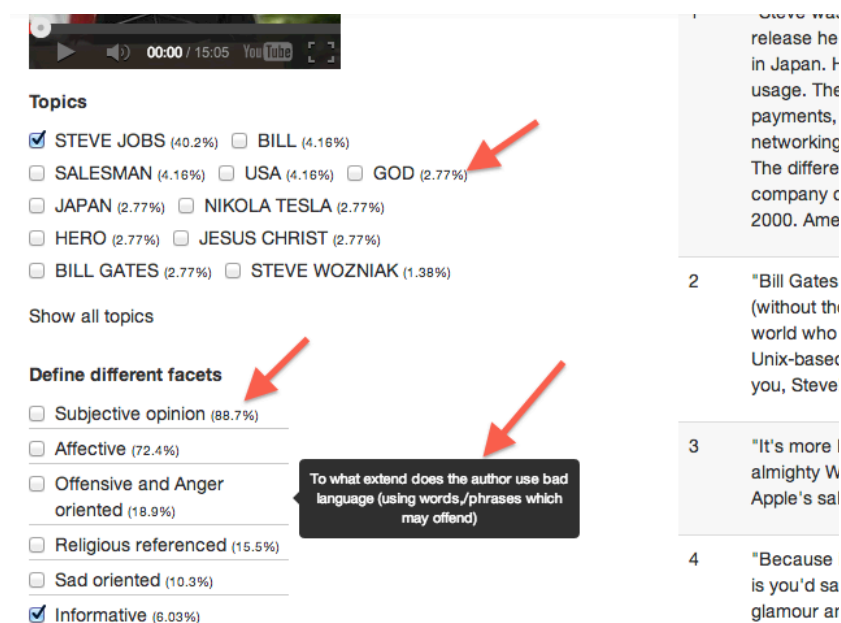


Figure 5.3: Screenshot 3 of the Web-based user interface of the framework. The system shows the frequency percentage of comments related to a facet and on-fly short overviews of a facet.

When a list of comments is shown based on a combination of facets, the system also shows a so-called on-fly short overview of all other possible facets for each comment (see Figure 5.2). Furthermore, with regard to suggestions received in our user study (Section 6.2, Study1), the service provides the conversation thread of each comment to the user, thus helping users to understand the context of a comment.

Additionally, in order to enable a user to quickly receive information about recommended facets (such as description, numbers of related comments, etc), the system also shows on-fly short overviews of such information besides facets (see Figure 5.3).

5.5 Summary

After having presented the AMOWA framework and its elements in various levels of abstraction, this chapter discusses the development of a Web-based interactive implementation of the AMOWA (Adaptive Moderation of Web Annotations) framework that allows users to work with the moderation model and framework using interaction metaphors. This interface enables end-users to moderate social media content based on their preferences and interests. Users provide feedback simultaneously by implicit means (using the faceted browser) or explicit means (voting). We discuss three implementations: *Usefulness Prediction Model*: we discuss a prototypical implementation of the prediction model for automatically predicting usefulness of UGC without receiving explicit or implicit users' feedback. *AMOWA-WS*: we discuss a prototypical implementation of a Web service of AMOWA which can be simply integrated as a plugin into any social media platform — or any platform which deals with UGC — and enables end-users to moderate content with regard to their personal interest or task in hand. *AMOWA-UI*: we discuss an implementation of a Web user interface of AMOWA. This user interface is a client-side implementation of the AMOWA-WS which allows users to access the Web service and to work with the moderation framework using interaction metaphors by exploring and selecting different combinations of facets. In order to give a better insight of functionalities of the AMOWA-UI, Figures 5.4 to 5.6 show samples of ranked results for top three facet combinations (evaluated in the Chapter 6 – Study 1). Figure 5.4 shows ranked

Multi-Faceted Adaptive Ranking Service for Social-Media Comments
LOGOUT
Instruction

Select preferred topic from the list

World War I Begins
Load

World War I - Treaty of Versailles

Topics

☒ GERMANY (16.0%)
☐ AUSTRIA (6.53%)
☐ HUNGARY (6.03%)
☐ VERSAILLES (6.03%)
☐ BELGIUM (5.52%)
☐ BRITAIN (5.02%)
☐ FRANCE (5.02%)
☐ USA (4.02%)
☐ EUROPE (3.01%)
☐ 1914 (2.01%)
☐ 1871 (2.01%)

Show all topics

Define different facets

☐ Subjective opinion (96.7%)
☐ Affective (77.1%)
☒ Informative (32.6%)
☐ Sad oriented (28.2%)
☐ Religious referenced (14.1%)
☐ Offensive and Anger oriented (11.9%)
☐ Chronological

[Reset selection](#)

You selected the following facet(s) and/or topic(s):
Topic: GERMANY and Informative

Available Results: 23

Relevant: 0 positive, 0 negative Interesting: 0 positive, 0 negative

#	Comment	Relevant	Interesting
1	"I wouldn't go as far as saying that most of the blame should fall on Germany, Germany kept their side of an alliance with Austria-hungary just like Britain kept to her side of an alliance with Belgium! WW1 occurred because the allies were cautious of Germany's expansion, it was "better now then later" to stop germany from expansion in Europe. That is completely hypercritical when England, France and Russia all had expansion desires of their own. WW1 was based on insecurity and Hypocrisy." (This comment is a reply. Click to show conversation)	<input type="button" value="Yes"/> <input type="button" value="No"/>	<input type="button" value="Yes"/> <input type="button" value="No"/>
2	The treaty of Versailles would have crippled Germany to this very day. Germany would not be the economic power it is today had they not broken the treaty. Germany acted within reason siding with Austria an ally and entering world war 1 they did nothing so out of the ordinary in the context of war to deserve such crippling economic punishment. Germany conducted the 1st world war in a gentlemanly fashion and even declared war before attacking. They let it be known that Belgium was the first front.	<input type="button" value="Yes"/> <input type="button" value="No"/>	<input type="button" value="Yes"/> <input type="button" value="No"/>
3	"germany didn't have 32 billion ...but the British , French did? when France and Belgium were devastated by the war...there was no devastation in Germany.So how is it that you think Germany should not pay to repair france and Belgium?Who should pay to repair France and Belgium?...no one?Just leave the whole area looking like the surface of the moon? Is that your solution?"	<input type="button" value="Yes"/> <input type="button" value="No"/>	<input type="button" value="Yes"/> <input type="button" value="No"/>

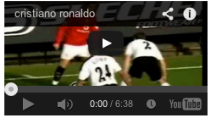
Figure 5.4: Screenshot 4 of the Web-based user interface of the framework. Comments are ranked (on a YouTube video, “World War I - Treaty of Versailles”) by selecting a combination of objective (“Informative” as a facet) and topic (“Germany” as a facet) facets.

comments by selecting a combination of objective and topic facets. Figure 5.5 shows ranked comments by selecting a combination of subjective and topic facets. Figure 5.6 shows ranked comments by selecting a combination of topic facets.

Multi-Faceted Adaptive Ranking Service for Social-Media Comments [LOGOUT](#) [Instruction](#)

Select preferred topic from the list

Cristiano Ronaldo



Topics

☒ CRISTIANO RONALDO (32.7%) ☐ STRIKER (9.50%)
☐ SWEDEN (5.28%) ☐ MADRID (4.57%)
☐ 2013 (2.81%) ☐ 2006 (2.46%)
☐ BARCELONA (1.78%) ☐ SOUTH KOREANS (1.76%)
☐ MANCHESTER (1.40%) ☐ OSCAR (1.05%)

Show all topics

Define different facets

☐ Subjective opinion (80.3%)
☐ Affective (72.1%)
☒ Offensive and Anger oriented (10.2%)

You selected the following facet(s) and/or topic(s):
 Offensive and Anger oriented and Topic: CRISTINO RONALDO

Available Results: 11
 Relevant: 0 positive, 0 negative Interesting: 0 positive, 0 negative


#	Comment	Relevant	Interesting
1	Messi is shit. Ronaldo beata messi	<input type="button" value="Yes"/> <input type="button" value="No"/>	<input type="button" value="Yes"/> <input type="button" value="No"/>
2	"FUCK YOU MESSI, RONALDO IS THE BEST <3"	<input type="button" value="Yes"/> <input type="button" value="No"/>	<input type="button" value="Yes"/> <input type="button" value="No"/>
3	And cristiano Ronaldo is much better on Real Madrid because he scored more goals and fuck you meshit	<input type="button" value="Yes"/> <input type="button" value="No"/>	<input type="button" value="Yes"/> <input type="button" value="No"/>
4	Im your biggest fan Ronaldo! You are really the best player. I put some videos of you on my channel please watch them! Pleeaaassee!!	<input type="button" value="Yes"/> <input type="button" value="No"/>	<input type="button" value="Yes"/> <input type="button" value="No"/>
5	"Are you crazy? Cristiano Ronaldo has 50 goals in 53 games, Ribery has 10 goals in 50 games. You are a bastard who does not know football."	<input type="button" value="Yes"/> <input type="button" value="No"/>	<input type="button" value="Yes"/> <input type="button" value="No"/>

Figure 5.5: Screenshot 5 of the Web-based user interface of the framework. Comments are ranked (on a YouTube video, “Cristiano Rolando”) by selecting a combination of subjective (“Offensive and Anger Oriented” as a facet) and topic (“Cristiano Rolando” as a facet) facets.

Multi-Faceted Adaptive Ranking Service for Social-Media Comments [LOGOUT](#) [Instruction](#)

Select preferred topic from the list

Hong Kong Returned to China 11



Topics

☒ HONG KONG (17.1%) ☐ CHINA (14.7%)
☒ ENGLAND (8.40%) ☐ BRITAIN (5.62%)
☐ USA (4.72%) ☐ COMMUNIST (3.22%)
☐ 1997 (1.72%) ☐ QUEEN (1.57%)
☐ SHANGHAI (1.12%) ☐ BEIJING (1.12%)

Show all topics

Define different facets

☐ Subjective opinion (95.9%)
☐ Affective (82.8%)
☐ Sad oriented (24.8%)
☐ Informative (23.8%)
☐ Offensive and Anger oriented (18.4%)
☐ Religious referenced (16.9%)
☐ Video timestamp (5%)

You selected the following facet(s) and/or topic(s):
 Topic: HONG KONG and Topic: ENGLAND

Available Results: 22
 Relevant: 0 positive, 0 negative Interesting: 0 positive, 0 negative

#	Comment	Relevant	Interesting
1	"Nobody in mainland doesn't know Hong Kong handover is by law, OK? In my opinion, only whose HongKongers who are unhappy with handover will say "UK abandoned HK". So, please make some sense, Hongkonger."	<input type="button" value="Yes"/> <input type="button" value="No"/>	<input type="button" value="Yes"/> <input type="button" value="No"/>
2	UK will take Hong Kong back and it will be free once again! Down with Communism!	<input type="button" value="Yes"/> <input type="button" value="No"/>	<input type="button" value="Yes"/> <input type="button" value="No"/>
3	"You are obviously a well balanced individual with a chip on both shoulders. My advice is to concentrate your rhetoric on China. You have convinced me that my nation is poor and unfortunate and yours is great. I might just move there in fact. On hang on a second maybe not. I enjoy freedom of speech. No censorship and I as a soldier, citizen and resident have definitely witnessed NO Tiananmen Square atrocities in the England in the last 38 years. Poor Hong Kong."	<input type="button" value="Yes"/> <input type="button" value="No"/>	<input type="button" value="Yes"/> <input type="button" value="No"/>
4	"If the UK had any intention of doing what the people of Hong Kong wanted, they would have gone to war for the city. They could have gotten the US to back them up and taken down the PRC and establish China as a true democratic people's republic."	<input type="button" value="Yes"/> <input type="button" value="No"/>	<input type="button" value="Yes"/> <input type="button" value="No"/>

Figure 5.6: Screenshot 6 of the Web-based user interface of the framework. Comments are ranked (on a YouTube video, “Handover 1997 Hong Kong”) by selecting a combination of topic (“Hong Kong” and “England”) facets.

Chapter 6

Experiments and Evaluation of Proposed Framework

6.1 Introduction

After having presented different types of developments of the AMOWA framework in the previous section in order to demonstrate the benefits of a proposed adaptive moderation framework we now turn to its evaluation through quantitative and qualitative evaluation of the framework and specifically we would like to answer the following questions:

1. How well does adaptive faceted ranking compare to the prevalent default method (reverse-chronological)?
2. Which type of facets perform best for ranking and what are effective strategies for selecting and building facets
3. How accurate is facet clustering?
4. Which topic identification algorithm is most appropriate for short user-generated content such as comments?

In order to answer these questions we set up three studies using our Web service and related user interface (AMOWA-WS and AMOWA-UI) and details implementation of the evaluation framework and results are under review for a publication [Momeni et al., 2014a]. First study utilizes a within-subjects design in order to compare the proposed framework to most prevalent default (reverse-chronological) ranking method. The results of this study are divided into two parts: (1) the quantitative assessment which measures the performance using Mean Average Precision (MAP). This measures the placement of interesting comments in the ranked results. (2) the subjective assessment which asks evaluators to answer questions regarding effectiveness, efficiency, and satisfaction of using such a system. Our second study evaluates the performance of clustering comments along different semantic facets and proposed semantic enrichment method. Our third study evaluates which topic-identification algorithm is most appropriate for short texts. This helps us to define an appropriate method for identification of topics, which can be used as a facet.

Dataset: The primary dataset used for this evaluation is described in Section 3.3. [Momeni et al., 2013a]. Briefly, this data set is compiled from real-world comments harvested from the popular social media platform YouTube. They are free-text comments on videos from a variety of users with different backgrounds and intentions. First, three broad entity types were selected: *event*, *person*, and *place*. For each entity type, a number of queries referring to historical topics in the 20th century were compiled. Examples of queries 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 queries. 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 proposed by the YouTube API. This enabled us to investigate the effectiveness of reverse-chronological ranking, which is a default setting. The distribution of videos and comments across the three entity types is shown in Table 6.1. Most comments belong to the event type while the fewest belong to place type; this skew can be explained by the data collection process: since videos with many comments or

	Event	Place	Person	Total
Videos	151	25	132	308
Comments	50,654	6,908	34,216	91,778

Table 6.1: Basic statistics of experimental data set (YouTube videos and comments).

many views are preferred, the more interesting videos are implicitly selected (which often happen to revolve around events instead of places).

6.2 Study 1: Effectiveness of Adaptive Faceted Ranking Strategies

We utilize a within-subjects design in order to:

1. compare the proposed framework to the standard reverse-chronological ordering approach.
2. investigate the effectiveness of strategies for the selection of different types of facets..

Note that we restrict our comparison of our proposed approach to the standard reverse-chronological ranking. We do not consider rankings generated by the crowd — users vote contributions of other users — as we observe that in the selected dataset, the average number of up-voted comments by the crowd for each video is at most four, which is not sufficiently representative for our comparison. This is not only an artifact of our data set, but a common problem in most social media platforms; users read comments but do not often use the voting function.

6.2.1 Participants

We recruited participants through two large universities (one in the US — University of Michigan — and one in Europe — University of Vienna), using computer and information science mailing lists. From the respondents, we randomly selected

36 (ages range from 20 to 57, median=29; 26 are students and 10 are other professionals such as administrative staff, lecturers, librarians, technical staff, etc). These participants all indicate that they frequently watch YouTube videos. Participants received a gift voucher for their efforts in evaluating the system.

6.2.2 Experimental Setup

Participants received training through an online instruction page (see Appendix2–Online Evaluation Instruction). After the training phase, they were asked to perform the following steps:

1. Use the prototype to select a title from a list of 30 videos (we restrict these to control for a reasonable, and approximately equal length and quality of video) and watch the video,
2. Use the prototype to retrieve a ranked list of the top 30 comments for a video based on reverse-chronological order.
3. 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.
4. Vote on each comment and each ranking condition. In the facet-based ranking, each comment is rated along two dimensions: *interestingness* and *relevance*. In the chronological ordering mode, only the *interestingness* is rated as *relevance* is a very ambiguous concept without selecting a particular facet (as we define a comment relevant when it is relevant to the facet selection of a user. For example, when a user selects subjectivity facet, the relevant comments are comments with higher subjectivity tone, but not necessarily relevant to the topic of the video.).

We restricted the size of the ranked list of comments to 30 in order to minimize judgment fatigue. For reverse-chronological rankings, users received the same set of ranked comments on the same video. 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, and,
- **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. Additionally, due to the fact that different combinations of facets result in various numbers of comments, we excluded all rankings which result in less than three comments (78 rankings out of 339 rankings). For example, if particular combinations for a video result in only one comment which is relevant, then MAP and $P@10$ are 1. This result would be quite unfair in comparison to other combinations returning 30 comments.

6.2.3 Results of Quantitative Assessment

The results of our analysis are shown in Table 6.2. The results of our analysis are shown in Table 6.2. As not all participants used all combinations, Table 6.2 also lists the number of rankings ($\#R$) for each combination and AF in table 6.2 shows performance of adaptive faceted ranking for any combination of facets (TF, OF, SF, or Any). Let us first consider our baseline, the reverse-chronological ranking (RC) — the effectiveness measures indicate that this ranking is at least somewhat effective. Approximately half of the comments retrieved are *interesting* to the users. In contrast, in the ranking of comments retrieved with our adaptive faceted ranking (AF) strategy, approximately every two out of three results are deemed interesting.

Table 6.2: Overview of effectiveness of adaptive faceted (AF) ranking strategies and the reverse-chronological ranking (RC) with respect to relevance and interestingness. Second column ”#R” shows number of rankings for each combination. The best performing facet combination is shown in grey.

Ranking Method (Facet Type)	#R	Interesting			Relavant		
		MAP	P@10	P@20	MAP	P@10	P@20
RC	51	0.46	0.48	0.53	<i>Not applicable</i>		
AF	214	0.71	0.57	0.52	0.80	0.70	0.61
AF (OF+TF)	26	0.88	0.80	0.77	0.90	0.75	0.75
AF (OF+SF)	12	0.51	0.40	0.47	0.65	0.65	0.67
AF (SF+TF)	35	0.77	0.57	0.58	0.89	0.68	0.68
AF (OF)	62	0.71	0.56	0.55	0.83	0.70	0.68
AF (TF)	49	0.72	0.63	0.63	0.80	0.77	0.76
AF (SF)	27	0.53	0.40	0.40	0.79	0.73	0.74

When considering different facet combinations, we observe that two combinations (SF, OF+SF) perform considerably worse than AF, two have a similar effectiveness (TF, OF) and two combinations (OF+TF, SF+TF) considerably outperform AF with respect to *interestingness* and MAP. The results are similar when considering *relevance* instead of *interestingness* (right part of Table 6.2). The strategy OF+TF is still the best performing one while the other strategy that exploit topic and subjective facets (SF+TF) performs only slightly worse. The worst performing strategy is now OF+SF. Thus, overall we have shown that semantic enrichment and a frequency-based facet selection schema (based on the Greedy Count algorithm) yield considerable improvements in terms of effectiveness compared to reverse-chronological ranking. We also conclude that topic facets are most important to a successful ranking strategy, both in terms of relevance as well as interestingness. Nevertheless, different combinations of other types of facets with topic facets perform slightly better and are more effective when comments do not explicitly represent a specific topic (see Section 6.4, Study 3).

With regard to the performance of faceted ranking concerning different types of facets for interesting votes (TF, SF, OF, and Combination of All), Table 6.2 shows that a faceted ranking based on combinations of topics and objective facets (OF+TF)

or combinations of topics and subjective facets (SF+TF) performs in an improved manner with regard to MAP. From .46 to .90 and from .46 to .77 respectively compared to reverse-chronological ranking. Furthermore, among all groups of facets, a faceted ranking based on combinations of topics and objective facets (OF+TF) performs in an improved manner (from .71 to .90) compared to other strategies and particularly compared to combinations of only topic facets (from .82 to .90) or combinations of only objective facets (from .70 to .90). Although a faceted ranking based on combinations of topics and subjective facets (SF+TF) performs in an improved manner (from .46 to .77) compared to reverse-chronological ranking, combinations of only subjective facets do not lead users to find interesting comments and performs almost slightly lower than reverse-chronological ranking (from .46 to .40). This is due to the fact that many users select the “Offensive and Anger Oriented” and “Self Reference” among subjective facets. In almost every case, this results in comments which are not interesting.

The proposed framework performs well for relevancy given a selected facet with a MAP (for all facets) at .83. Table 6.2 also shows that a faceted ranking based on combinations of topics and objective facets (TF+OF) or the combination of topics and subjective facets (TF+SF) performs better with regard to MAPs, .92 and .91 respectively) compared to combinations of only one type of facets such as only subjective or only objective (.67 and .78 respectively). However, combinations of only topic-related facets also perform with high reliability (.90).

These numbers are useful but do not tell us whether adaptive faceted ranking statistically outperforms reverse-chronological ranking. Therefore, for all faceted ranking results which have an equal number of comments to chronological ranking (30 comments) by the same users on the same videos (105 rankings from 339 rankings in total), we create two sets: (1) all positive and negative interesting votes collected for faceted ranking results and (2) all positive and negative interesting votes collected for chronological ranking results. We add ranking results related to the same videos by the same users in the same order in each set and then apply Pearson’s Chi-squared tests for different types of faceted ranking compared to reverse-chronological ranking. Results of this study in Table 6.3 indicate that majority of adaptive faceted rankings statistically significantly outperforms reverse-chronological ranking. This

Table 6.3: Adaptive faceted ranking has statically significant difference compare to reverse-chronological ranking. The bolded mean values point out considerable positive differences between positive and negative votes. The star next to the X2 means that there is evidence ($p < 0.001$) that the two predicted samples come from different distributions.

Facets' Combinations	X2	Mean- Positive- Votes	Mean- Negative- Votes	STD
<i>All</i>	20.66*	16.18	13.91	7.81
<i>SF+TF</i>	791.26*	20.70	09.30	6.7
<i>OF+TF</i>	845.31*	24.50	05.50	4.27
<i>SF+OF</i>	123.61*	16	14	5
<i>TF</i>	874.13*	18.45	11.54	7.92
<i>SF</i>	385.63*	11	19	6.53
<i>OF</i>	301.12*	19.02	10.97	6.75

shows that adaptive faceted ranking increases users' ability to read comments they wish to see. All facet combinations selections which result in less than 30 comments are excluded from this test. Furthermore, Table 6.3 shows mean and standard deviation values of positive and negative voted on comments for different facets. These results confirms that the combination of Objective and Topic facets is more effective compare to combinations of other types of facets.

These results indicate that adaptive faceted ranking that is supported by the semantic enrichment performs better in comparison to reverse-chronological ranking. Furthermore, our results demonstrate that multi-faceted rankings (combinations of different facets) perform better in comparison to faceted rankings using only one type of facets (such as subjective facets or topics alone). In addition objective facets are desirable (over subjective facets or topics alone) and may argue for additional facets of this type. Nevertheless, it is important to note that the adaptive ranking based on topics also performs well. However, the adaptive multi-faceted ranking performs slightly better and is more effective when many comments do not explicitly contain a specific topic (see Section 6.4, Study3).

Having collected for each comment in the selected dataset an "Interesting" and a "Relevant" vote, Figure 6.1 shows that what the majority of users have selected as interesting is also relevant to what they have selected as facets. However, what

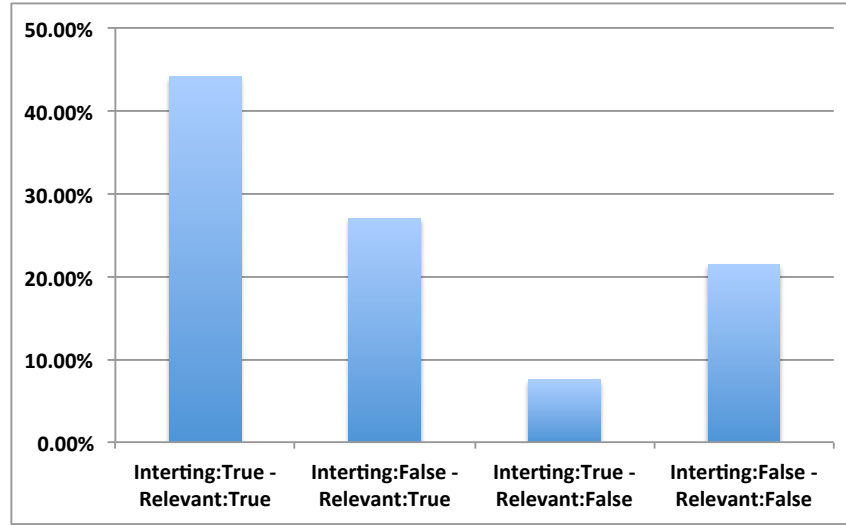


Figure 6.1: Percentages of comments with various combination of “Interesting” and “Relevant” votes, which shows comment which is relevant to selected facts by a user is not necessarily interesting for the same user and vice versa

they have selected as relevant is not necessarily interesting. These results confirm our assumptions that two dimensions of usefulness should be take into consideration separately.

6.2.4 Results of Subjective Assessment

Our proposed system design hypothesizes that an adaptive faceted ranking can help users manage the overwhelming data in their comment feeds. To get a broad sense of whether the proposed approach is subjectively better for browsing an overwhelming comment feed than standard reverse-chronological ranking, participants ask to report subjective ratings on a 5-point Likert scale for the proposed framework (with 1 meaning “strongly disagree” and 5 meaning “strongly agree”) by answering ten subjective statements:

1. I think that I would like to use this system frequently.
2. I found the various functions in this system were well integrated.

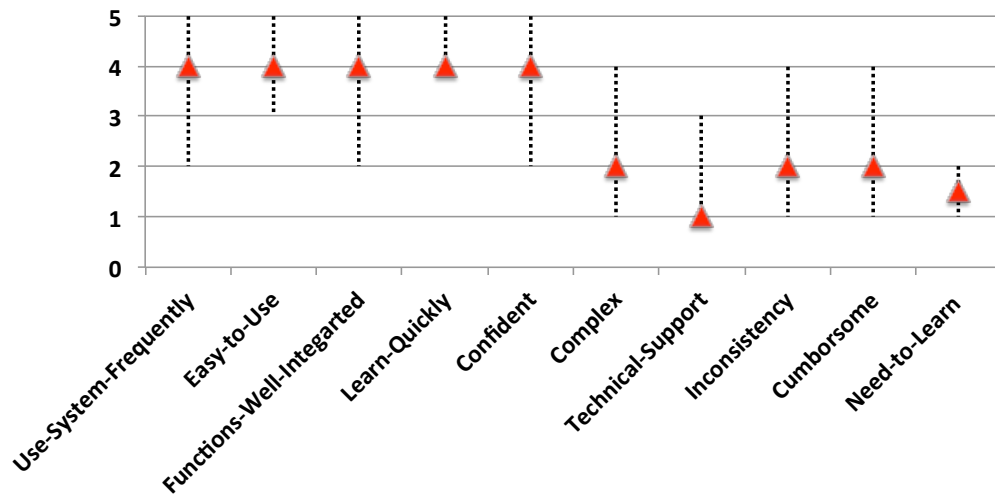


Figure 6.2: Overview of the subjective evaluation

3. I thought there was too much inconsistency in this system.
4. I found the system unnecessarily complex.
5. I thought the system was easy to use.
6. I think that I would need the support of a technical person to be able to use this system.
7. I found the system very cumbersome to use.
8. I felt very confident using the system.
9. I needed to learn a lot of things before I could get going with this system.
10. I would imagine that most people would learn to use this system very quickly.

Our evaluation measures subjective outcomes outlined by Hearst [Hearst, 2009] which are important constructs for the evaluation of interfaces such as effectiveness (related to statements 1 to 4), satisfaction (related to statements 5 to 9), and efficiency (related to statement 10). Figure 6.2 shows an overview of the subjective evaluation.

Effectiveness and satisfaction are the two greatest benefits from using adaptive faceted ranking. The majority of users reported that they would use this system frequently and felt confident when using the system. Furthermore, end-users found the faceted ranking system easy to use, the various functions in the system are well-integrated, and there was not too much inconsistency in the system. One participant commented: *“I like the structured approach to finding comments that are relevant to my interest,”* while another participant said: *“I think it is a really good idea to be able to filter through the comments, especially if there are lots of them.”* Furthermore, efficiency is considered one of the positive benefits from using faceted ranking. One participant wrote: *“I liked that the system works quickly and that it allows the user to combine Topics and Facets according to their needs and interests.”* Also, another user explicitly reported, *“I found that the facets were more interesting to filter instead of the topics.”*

Moreover, most users agree that most people can learn to use this system very quickly. Besides, there is strong agreement that the support of a technical person to be able to use this system is not necessary. Nevertheless, some users (2 users) indicate that the system is cumbersome, but on further exploration we found they were actually referring to the evaluation system (which required that they vote on comments to complete the task). For example, one user said *“It is a little bit cumbersome, because so many things must be voted.”*

When asked about the worst part of the user interface, one user explained *“I think the function is great, however, the way the topics, facets and comments displayed is not that friendly.”* Another user explained *“the comments weren’t that relevant to some facets. especially for the offensive one. take into consideration of slang and how people express their special awes [sic].”* Finally, participants also provided us with interesting suggestions which will be taken into consideration in future work. Three users suggested *“I’d like the system to be able to pick up advertising comments where people just ask you to go check out their channel, and have that as a separate facet.”* Another user explained that *“[I] would prefer to use this system to filter out things [I] don’t want to see– like advertisements or flame wars – rather than looking for something specific in comments.”* Finally, it is interesting to note that the evaluation was performed in two rounds. Based on some suggestions received

in the first round, we improved the development of the user interface. For example, having received a suggestion from three users in the first round such as “*it would be nice to be able to follow conversations in the comments using this tool*”, we added the conversation thread of comments for selected comments in the study.

6.3 Study 2: Faceted Extraction and Ranking Performance

This study evaluates the effectiveness of clustering comments along different semantic facets.

6.3.1 Experimental Setup

Crowdsourcing-based evaluation In order to evaluate the accuracy of our semantic enrichment and facet clustering approaches, we created a ground truth dataset by annotating a subset of the comments. Specifically, for ten randomly selected videos by users from Study 1 (Section 6.2), 100 comments were randomly chosen (thus in total 1,000 comments). The comments were annotated with respect to *Subjectivity*, *Affect*, *Offensiveness*, *Video Timestamp*, *Sadness*, *Anger*, etc. For the annotation process, we relied on crowdsourcing and employed workers via Amazon’s Mechanical Turk¹ 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. Workers not performing satisfactorily on this question were excluded from further participation. Additionally, workers had to provide binary judgments for each of the eight facets listed above. Thus, overall nine questions (including the honey pot question) had to be answered by each worker for each comment.

Having collected three judgments per comment, we first determine the inter-rater

¹<https://www.mturk.com/mturk/>

Table 6.4: Judges' inter-agreement for each proposed facet based on Fleiss' Kappa.

Facet	Type	Fleiss' Kappa	%Comments
<i>Subjective</i>	SF	0.67	%96
<i>Affective</i>	SF	0.75	%79
<i>Religious Referenced</i>	OF	0.76	%10
<i>Video Timestamp</i>	OF	0.97	%2
<i>Offensive and Angry</i>	SF	0.79	%17
<i>Informative</i>	OF	0.67	%62
<i>Sad</i>	SF	0.78	%14

agreement for each facet based on Fleiss' Kappa. The results are shown in Table 6.4. The agreement is close to perfect for the *Video Timestamp* facet, which 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. This table also shows percentages of comments for each facet. Examples of comments labelled with these facets (with high and low agreement) are shown in Table 6.5.

Overall, we consider a Kappa above .65 for all but one facet as substantial agreement between the raters. Comments are labeled along different features based on majority agreement (when two out of three coders agreed). For example, when two or three coders agree that a comment is "Subjective", then we labeled this comment as "Subjective".

We measure the effectiveness of our clustering approach by Precision, Recall, and F1 for each facet.

6.3.2 Results

The effectiveness of our clustering and facet extraction approach are shown in Table 6.6. 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 clustering and extracting facets related to the facets *Subjectivity*, *Affective*, and *Video TimeStamp*. However, the clustering of facets *Informativeness*, *Religious Referenced*, *Sad*, *Offensive and Angry* proves to be more difficult. This difference in clustering

Table 6.5: Examples of comments that achieved full vs. moderate annotator agreement. The three facets shown are those with the lowest inter-annotator agreement.

Facet	Full agreement	Moderate agreement
<i>Subjective</i>	<i>“One of the greatest speeches i’ve ever eared..GRAZIE Steve”</i>	<i>‘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”</i>
<i>Informative</i>	<i>“Austria and Hungry was a major ally of Germany. They helped the Germans annihilate the russian army.”</i>	<i>“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”</i>
<i>Affective</i>	<i>“Such an awful thing to happen to such a peaceful and talented man :(R.I.P John Lennon.”</i>	<i>“If there was one thing everyone involved in the war could agree on, it’s that they did not like Versailles.”</i>

effectiveness is a reflection of the difficulties our human annotators had in this task.

After a manual inspection of the results returned for the facet, *Self-reference* in the test phase of manual coding, it was determined that coders had the highest disagreement on coding this facet. Thus, it was removed from further consideration due to its noisy nature.

6.4 Study 3: Comparison of Topic Detection Algorithms

In order to explore which topic-identification algorithm is most appropriate for short texts, the performance of different topic identification algorithms is experimented with.

Table 6.6: Overview of clustering performance across all facets ordered by their accuracy.

Facet	Type	Precision	Recall	F1
<i>Video Timestamp</i>	OF	0.91	0.91	0.91
<i>Subjectivity</i>	SF	0.95	0.98	0.97
<i>Sad</i>	SF	0.78	0.65	0.71
<i>Religious referenced</i>	OF	0.58	0.88	0.70
<i>Offensive and Angry</i>	SF	0.66	0.90	0.76
<i>Affective</i>	SF	0.92	0.91	0.92
<i>Informative</i>	OF	0.86	0.61	0.71

6.4.1 Algorithms

We empirically evaluate three approaches:

1. *TF-IDF based on Unigrams*: The unigrams with the highest TF-IDF score are utilized. This approach does not require external resources nor is it 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 employ the semantic enrichment service GATE². In order to ensure a very high accuracy and to disambiguate entities, we apply a simple method by calculating the similarity scores among the letters of Named Entities in the context of all comments on a video, comparing all highly similar entities manually, and creating a list of ambiguity Named Entities – since our goal is not to evaluate the effectiveness of a particular entity detection approach, but to evaluate the ability of NEs in general to act as proxies for topics.
3. *Topic Modeling (LDA)*: Lastly, we experiment with statistical topic modeling, in particular Latent Dirichlet Allocation (LDA) [Blei et al., 2003]. An LDA model is trained by aggregating all comments for a specific video and inferring the topic distribution from this aggregate (the following standard

²<https://gate.ac.uk/>

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.

6.4.2 Experimental Setup

From each approach, we utilize the highest scoring topic label. A concrete example of extracted topic labels is shown in Table 6.7 for four comments.

Table 6.7: Topic label examples. Bolded items shows topic labels with highest agreement among coders

Comment	TF-ID-based	Entity-based	LDA
<i>“it was a white Fiat, and it was later found with the owner burnt to death inside. I believe his name was Anderson.”</i>	Owner	ANDERSON	White
<i>“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.”</i>	Board	TONY HAWK	Matter
<i>“Just legalize everything. By making these drugs illegal, you’re giving the criminals their business. It’s almost the equivalent of living off of a government paycheck, which is why the criminals loved Prohibition.”</i>	Criminals	Not applicable	Government
<i>“Sometime i lie. sometime i speak of the truth. Every lie and truth has a plan and meaning. ”</i>	Lie	Not applicable	War

Crowdsourcing-based evaluation In order to evaluate the three approaches, we randomly pick 1,000 of our available comments and present the comments along with their extracted topic labels to Amazon Mechanical Turk workers. Each worker is 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 crowd sourcing-based evaluation in Study 2 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 generate three topic recommendations (one topic by each of the 3 approaches). Similar to previous work, we collect three worker judgments per a topic label and comment pair.

6.4.3 Results

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, the TF-IDF based topic labeling outperforms LDA with a relevant label for 67% of the comments. We find that the LDA Analysis generally does not provide meaningful topic terms (Table 6.7 shows examples of such terms). Also, providing interpretable descriptions for topic models is a difficult problem. Besides, even “optimal” models may not be consistent with reader preferences [Boyd-Graber et al., 2009]. This results shows that for the problem of extracting a relevant label from a comment, the Entity-Based approach performs better than the investigated alternatives for those comments with Named Entities occurring.

In the next step, we employ a logistic regression analysis to identify the likelihood of a binary output (similar to the method used by [Bernstein et al., 2010]) and we measure with odds-ratio and coefficient ranks. Table 6.8 shows detailed results of this study. The coefficient ranks demonstrate that the NE-based approach outperforms the other two. More precisely, the coefficient rank for the NE approach indicates that using the Entity-Based algorithm has positive correlation with providing a relevant topic for a comment. However, using Topic Modeling algorithms has negative correlation with providing a relevant topic for a comment.

Table 6.8: Coefficients and Odds-Ratios of different topic labeling approaches evaluated on 1000 comments.

Algorithm	Odds-Ratio	Coefficient
<i>TF-IDF based on Unigrams</i>	1.19	0.17
<i>Entity-based</i>	11.79	2.46
<i>LDA-based</i>	0.19	-1.64

6.5 Discussion

Many available approaches employ topic-based browsing so that users are able to more efficiently browse their feed and hide irrelevant content based on users' preferences. However, comments are often very brief and topics discussed alongside comments are often very noisy. Furthermore, comments which are clustered according to an explicit facet only based on their topics result in a single imperfect faceted ranking. This ranking does not enable users to rank comments with regard to other potentially useful facets. It is also important that systems help individuals adapt ranking based on the particular objective which the user happens to have in mind. Accordingly, we propose that an adaptive, personalized ranking of comments is desirable.

Our experimental results indicate that when semantic enrichment supports adaptive faceted ranking, performance is better in comparison to reverse-chronological ranking. Furthermore, our results demonstrate that objective facets are desirable (over subjective facets or only topics) and this may indicate that additional facets of this type are required. With regard to the performance of faceted ranking concerning different types of facets (Topic facets, Objective facets, Subjective facets, and Combination of All), results show that combinations of topic and objective facets perform in an improved manner compared to other combinations. Nevertheless, it is important to note that the adaptive ranking based on topics also performs well. However, the adaptive multi-faceted ranking performs slightly better and is more effective when many comments do not explicitly present a specific topic.

Moreover, usefulness for an individual confounds two aspects, *relevancy* to what they are looking for and *personal interest*, that should be treated separately. With

regard to these two types of votes, we discover that for the majority of users the set of comments which is voted as interesting is not equal to the set of comments which is voted as relevant. Therefore, for capturing both these dimensions, we suggest that the proposed faceted ranking framework should give users the chance to provide two explicit votes: *Relevant* and *Interesting*. Relevance votes capture the context or what a user is looking for, while interesting votes capture a user's personal interest.

Chapter 7

Conclusions

7.1 Discussion and Experimental Results

This work proposes an alternative, *automated* support for the *multi-faceted adaptive moderation* of user-generated content on the Web. The proposed approach is a semi-supervised learning approach for adaptive moderation of social media content with regard to the preferences of each individual user. It is influenced by past work on multi-faceted search [Koren et al., 2008], active learning, and topic identification. The adaptive moderation framework is built on the requirements derived from an analysis of current approaches in assessment and ranking methods of user-generated content in various application domains. In the realization of this framework, we derive a concrete application programming interface and a concrete representation of the prototypical Web-based user interface. Furthermore, as a part of the framework’s requirements, we try to better understand the characteristics of useful user-generated content and their prevalence patterns across different social media platforms.

We summarize the contributions and related experimental results related to each contribution as follows:

- ***C1: We carried out a comprehensive state-of-the-art analysis of the existing methods and approaches for assessment and ranking of UGC.*** The results of a systematic review of approaches for assessing and

ranking UGC with regard to three aspects are presented: “applied methods”, “values which are expected to be maximized”, and “application domains”.

With regard to “applied methods”, it is observed that many platforms use the crowd-based approach as a prevalent default approach. However, most of the proposed assessment and ranking approaches based on community-based assessment and ranking utilize machine-based methods for assessment of UGC. Examining machine-based methods more closely reveals that some machine-based assessment approaches include crowd judgments on the content in order to create a ground truth and some completely exclude crowd. On the other hand, many machine-based approaches exclude crowd for three reasons: (1) Different biases of crowd-based approaches such as “imbalance voting”, “winner circle”, “early bird”, etc.[Liu et al., 2007]. (2) A lack of an explicit definition of value which may be requested by the crowd to assess some application domains. For example, most approaches related to assessment of deceptive and spam product reviews exclude crowd-based judgments due to the fact that no platforms or domains have asked the crowd for deceptive judgments. (3) Human judgments can not be as precise as machine-based judgments in the case of some application domains and values (such as identification of truthful product reviews [Ott et al., 2012]).

Furthermore, it is observed that there is less consideration of the personalized and adaptive definition of the value of the individual user and most of the available approaches rely on particular sources of ground truth and do not enable users to make personal assessments of a particular value. This means there are few approaches which aim to accommodate individual differences in the assessment and ranking of UGC. For example, most of the work on identification of helpfulness of product reviews creates and develops prediction models based on a set of majority-agreement labeled reviews. However, helpfulness is a subjective concept that can vary for different individual users. Therefore, it is important that systems help individuals to make personal assessments of a particular value. Moreover, most of the available assessment and single-user ranking approaches focus on maximizing different values mainly for two application domains, postings in micro-blogging

platforms and postings in forums. These approaches mainly focus on creating interfaces that enable users to more efficiently browse their feed by providing a browsable access to all content in a user's feed and allowing the user to more adaptively find content related to her interests. At the back-end of these interfaces, there are two types of methods: (1) an algorithm which concurrently exploits the patterns of assessment and ranking settings by users to minimize the cost of changing settings for other users. This method leverages ideas from collaborative filtering and recommender systems [Lampe et al., 2007, Hong et al., 2012, Uysal and Croft, 2011]. (2) an algorithm which extracts a set of computational information cues from a set of content that can be used in the user interface, such as extracting a set of topics [Bernstein et al., 2010]. This means grouping a user's feed into consistent clusters of related concepts. However, these approaches are sometimes considered to be computationally costly, noisy, and require too many adjustments to work effectively across a wide range of users due to the fact that users try to post short texts in order to save space. Therefore, alternative approaches which take into consideration the semantic of the content or leverage the users' social networks for providing high quality rankings are required [Burgess et al., 2013].

With regard to "different values" which are expected to be maximized, some features have high impact for assessment of a particular value based on our analysis of features. Therefore, for maximizing some values, systems should take into consideration an easier way to build influential features at the design phase. For example, when maximizing value related to usefulness for comments on online media objects (such as YouTube videos), the system should encourage users and provide them with the opportunity to define references for enriching the texts of comments semantically [Momeni et al., 2013a]. Also, value related to credibility should take authors' profile pages into consideration [Morris et al., 2012]. Instead, many approaches which aim to maximize quality generally apply a crowd-based method. Besides, it is observed that many approaches related to assessment of the relevancy of UGC employ unsupervised learning approaches due to the fact that relevancy is influenced by

textual features and, therefore, applying unsupervised text clustering methods is effective for maximizing this value. Many approaches which aim to maximize helpfulness are mainly discussed in the domain of the product review and use crowd-judgments as the ground-truth to build their prediction model. However, the use of crowd for this value is a matter which provokes discussion. Similar to helpfulness, spam and deception are also principally discussed in the domain of the product review and how they differ in that they tend to exclude crowd-judgments. Approaches which are principally related to the assessment of popularity develop their identification and prediction models based on votes and ratings of crowd (in the case of Tweeter, re-tweet).

With regard to “application domains”, a more detailed examination leads to the discovery of many proposed machine-based assessment approaches in the Q&A domain which utilize semi-supervised learning approaches such as co-training or mutually reinforcing approaches. This is due to the fact that the interconnectedness and interdependency between the three sets of entities in Q&A (questions, answers and authors) is high. In addition, most of the available approaches focus on maximizing different values for micro-blogging platforms. This may be due to the very simple and structured characteristics of these platforms. Yet, there are fewer approaches to maximize important values for many application domains such as UGC on online media sharing platforms (e.g., YouTube, Flickr).

Based on these observations, we conclude that there are number of challenges which should be taken into consideration:

- How can the conceptual gap between crowd-based and machine-based approaches for optimizing assessment and ranking of the UGC be bridged? This challenge triggers many technical challenges which include: how can we develop algorithms and methods for preventing biases of the crowd? how can we take advantage of semi-supervised learning such as active learning for efficient integration of the crowd into machine-based approaches? or how can we utilize crowd to optimize the process of labeling large amounts of unlabeled UGC and improve the accuracy of hard machine-based judgments?

- How can we help people make personal assessments of a particular value rather than rely on particular sources as authorities for ground truth or minimize the amount of controversial assertions of value among users?
- ***C2 & C3: We gather a dataset of comments on online multimedia objects and we conduct different experiments for identification of the characteristics of useful comments.*** We conducted an analysis of user-generated comments on media objects of different social media platforms to examine the characteristics of useful comments and identify the important key features of comments for inferring usefulness. In order to achieve these goals, we analyzed three different sets of features: “Text Statistics and Syntactic”, “Semantic and Topical”, and “User and Social” features.

Our experimental findings show that “Semantic and Topical” features play important roles for inferring the usefulness of comments. For characterizing and inferring the usefulness of comments, a few relatively straightforward features can also be used. Comments are more likely to be inferred as useful when they contain a higher number of references, a higher number of Name Entities, and a lower self-reference and affective process (lower sentiment polarity, lower subjectivity tone, swear score, etc). Therefore, we suggest that a commenting system should urge users to define references [Haslhofer et al., 2013] by adding unambiguous concept references verified by users to social media comments. This in turn has a positive impact on the usefulness of comments.

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

An analysis of the important features among different topics (place, person,

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

With regard to the analysis of the prevalence of useful comments, our findings indicate that prevalence is influenced by the commenting culture of platforms and the different dimensions of topics of media objects. Also, the time period of topics has a slight influence on the usefulness prevalence. The nearer the time period of a topic is to the present time, the lower the prevalence of useful comments is. Moreover, the polarization of topics among commenters has a negative impact on the prevalence of usefulness. This means that for highly polarized topics the prevalence of useful comments decreases. Finally, we find that different platforms (Flickr and YouTube) lead to different prevalences of useful comments. For all entity types of topics (place, person, and event), the prevalence of useful comments on Flickr is higher than that on YouTube, which contains many more non-useful comments.

These results demonstrate that different platforms (Flickr and YouTube) lead to different usefulness classification results and the influence of features may vary according to the commenting cultures of platforms. Therefore, for a more accurate moderation of useful comments, a classification model should be trained with regard to the commenting culture of a platform and the preferences of the requester of moderation. Moreover, usefulness for an individual confounds two aspects, the *relevancy* of what she is looking for and her *per-*

sonal interest in what she attracts her attention. These aspects should be treated separately. Therefore, for capturing both these dimensions, the proposed faceted ranking framework provides the user with the chance to provide two explicit votes: *Relevant* and *Interesting*. Relevance votes should capture the context or what the user is looking for and interesting votes should capture a user’s personal interest.

- ***C4: We draw a number of conclusions and requirements for an adaptive moderation framework.*** Taking into account the results of analyzing the state-of-the art (available approaches related to ranking and assessing UGC) and our experiments regarding “Usefulness” identification, a number of fundamental problems appear:

1. *Biases of judgements by crowd:* The wisdom-of-the-crowd approach simply allows all users to vote on (thumbs up or down, stars, etc.) or rate comments. However, this approach avoids an explicit definition of usefulness. Furthermore, crowd-based voting is influenced by a number of biases such as “imbalance voting”, “winner circle” (e.g., a “rich get richer” phenomenon), “early bird” etc. that may distort accuracy [Liu et al., 2007].
2. *Removal of control from end-users:* Many approaches which are trained to rank comments are based on a set of majority-agreed labeled comments [Momeni et al., 2013a, Siersdorfer et al., 2010]. 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 requesters to personalize ranking based on their preferences.
3. *Complexity of usefulness:* Automatic moderation of comments by “usefulness” is generally complex, mainly due to the subjective nature of “useful”. In addition, even human raters find it difficult to agree on the usefulness of comments [Momeni et al., 2013a]. Moreover, usefulness for an individual confounds and blends together two aspects, *relevancy* to what they are looking for and *personal interest*. These should be treated separately. As a result, it is important that systems take into consideration both these dimensions of usefulness and help individuals adapt

ranking based on the particular objective which the user happens to have in mind.

4. *Comment as short texts:* Comments are often short and they are as fast for users to preview as to read completely. Often, they have no intermediate representation like a headline that can be used for searching and browsing news topic interfaces. Many topic-based browsing approaches propose strategies for extracting topics by enriching the semantics of an individual post. These approaches address both content and context by learning user preferences and hiding irrelevant content. However, comments are often very brief and topics discussed alongside comments are very noisy. Furthermore, as comments have multiple explicit dimensions (such as language tone, physiological aspects, etc), grouping them exclusively based on topic results in a single imperfect faceted ranking does not enable users to rank comments with regard to other potentially useful facets. Therefore, a system which combines higher level features alongside topic classification is desirable.
5. *Various cultures for generating content in different platforms:* Previous work [Momeni et al., 2013a] demonstrates that different platforms (Flickr and YouTube) lead to different usefulness classification results and the influence of features may vary according to the commenting cultures of platforms. Furthermore, with regard to the analysis of the prevalence of useful comments presented by Momeni et al [Momeni et al., 2013a], findings indicate that prevalence is influenced both by the commenting cultures of platforms and the different dimensions of topics of media objects. The time period of topics has slight influence on the usefulness prevalence. Therefore, for a more accurate classification of useful comments, a classification model should be trained with regard to the commenting cultures of platforms and media objects.

The issues discussed have led us to conclude that the following aspects are required for the development of an adequate adaptive moderation framework:

- A number of strategies for extracting novel facets and topics from com-

ments that operationalize the complex dimensions of usefulness. These strategies also define the benefits of combining different types of facets (such as a facet related to topic, subjectivity, etc) for providing end-users with access to interesting or relevant comments.

- An interactive framework for leveraging these facets to directly enable end-users to rank comments adaptively based on their preferences and interests with regard to the commenting culture of a platform.
- A possibility for users to provide feedback simultaneously by implicit means (using the faceted browser) or explicit means (voting). Both of these can be utilized to build user models and improve the automated moderation processes.
- A possibility to assess usefulness without users' feedback. While it is preferred that the feedback is provided by the user, it is helpful to begin with a "baseline" assessment of usefulness that is independent of the user.

- ***C5 & C6 : We further anticipate implementations of the proposed framework and we develop a Web-based interactive implementation of AMOWA (Adaptive Moderation of Web Annotations)*** by building a concrete basis for implementations of our model and specifying a generic application programming interface that covers static and dynamic aspects of the proposed framework. This generic interface specification allows for the implementation of the envisioned moderation framework in a number of application domains.

- ***C7: We demonstrate the benefits of a proposed adaptive moderation approach.*** In order to demonstrate the benefits of a proposed adaptive moderation approach for providing end-users with access to a useful, through to a quantitative and qualitative evaluation of the framework, we try to answer the following questions:

1. How well does adaptive faceted ranking compare to the most prevalent default (reverse-chronological) ranking methods used by different platforms?

2. What facets perform best for ranking and allow users to find interesting comments based on their preferences?
3. How accurate is facet clustering?
4. Which topic identification algorithm is most appropriate for short user-generated content such as comments?

In order to answer these questions, we set up three studies using our Web service and related user interface (AMOWA-WS and AMOWA-UI). Our first study utilizes a within-subjects design in order to compare the proposed framework to the commonly used default (reverse-chronological) ranking method. The results of this study are divided into two parts: (1) The quantitative assessment which measures the performance using Mean Average Precision (MAP). This measures the placement of interesting comments in the ranked results. (2) The subjective assessment which asks evaluators to answer questions regarding the effectiveness, efficiency, and satisfaction of such a system. Our second study evaluates the performance of clustering comments along different semantic facets and proposed semantic enrichment methods. Our third study evaluates which topic-identification algorithm is most appropriate for short texts. This study helps us to define an appropriate method for identifying topics which can be used as facets.

Also, we show that adaptive faceted ranking outperformed reverse-chronological ranking. However, we believe that chronological ranking is still useful for users with regard to the particular task in their minds. Therefore, we suggest that chronological ordering may be designed and developed as one of the facets to be suggested to users.

Our experimental results indicate that when semantic enrichment supports adaptive faceted ranking, performance is better in comparison to reverse-chronological ranking. Furthermore, our results demonstrate that objective facets are desirable (compared to subjective facets or only topics). This may be the reason for additional facets of this type. With regard to the performance of faceted ranking concerning different types of facets (Topic facets, Objective facets, Subjective facets, and Combination of All), our results show

that combinations of topics and objective facets perform in an improved manner compared to other combinations.

With regard to the topic identification algorithm, which is most appropriate for short texts, we find that the Entity-Based algorithm, outperforms other topic identification algorithms such as TF-IDF based on unigrams or LDA. Furthermore, we found that the LDA Analysis generally does not provide meaningful topic terms (generally identifying related terms rather than meaningful topics). In addition, providing interpretable descriptions for topic models is a difficult problem. Besides, even “optimal” models may not be consistent with reader preferences [Boyd-Graber et al., 2009]. Therefore, our results suggest that the topic identification of comments benefits from extracting comments’ Named Entities.

7.2 Conclusions and Future Directions

7.2.1 Limitation and Future Work

The result of our user study is limited to a comparison of adaptive faceted ranking with reverse-chronological ranking. However, it would be interesting to compare the faceted ranking with another common default ranking provided by the crowd — users vote the contributions of other users. However, we observe that in the selected dataset the average number of positively voted comments by the crowd for each video is maximum 3-4 comments. This is not sufficiently representative for our comparison and would therefore be a very interesting aspect to be studied in the future.

Although extracting useful comments from YouTube (only 8% of 3,500 comment) is very challenging as shown by Momeni et al [Momeni et al., 2013a], we show how our proposed ranking framework helps to extract a higher number of interesting comments with regard to users’ interests. Having selected one of the challenging platforms for this evaluation, we believe that the framework we propose will be adaptive and integrate with other platforms.

This work principally focuses on examining different strategies for semantic extraction of facets and adaptive ranking, by primarily using the explicit feedback to evaluate the performance of the proposed adaptive faceted ranking strategies. However, we will explore the personalized ordering of facets and ranking strategies in future work, thus enabling the personalization of the faceted ranking to a given user profile which is generated by the user modeling, and perhaps improving the results of faceted ranking.

Most of the available Named Entity Recognition tools provide ambiguity entities and Named Entity extraction is still not optimal. This is a limitation for extracting topics using Named Entities. Consequently, some fine tuning is required when named entities are used as a topic proxy.

Finally, we assume different designs and orderings of facets also have a significant impact on the results achieved. This may also be another interesting path to develop the project.

7.2.2 Summary and Conclusions

Considering the results of analyzing the state-of-the art (available approaches related to ranking and assessing user-generated content) and results of our experiments regarding the useful comment moderation, we observe a number of fundamental problems. These are biases of judgements by the wisdom-of-the-crowd approach, removal of control from end-users by many machine-based approaches which are based on a set of majority-agreed labeled comments, complexity of usefulness mainly due to the subjective nature of “useful” (in addition, even human raters find it difficult to agree on the usefulness of comments), UGC as short texts which have no intermediate representation like a headline that can be used for searching and news topic browsing interfaces, and various cultures for generating content in different platforms.

With regard to these observations, in this work we describe a novel adaptive faceted moderation framework for user generated content on the Web, which clusters adaptively each element of a comment along multiple explicit semantic facets (e.g., sub-

jectivity, informative, and topics) and then allows end-users to explore different clusters and select combinations of facets in order to moderate and rank comments that match their interests. The proposed framework comprises four main components: “*Semantic Enrichment Comments*”, are often short and do not explicitly feature facets which describe their content. The proposed moderation framework first enriches each element of comment along various semantic facts and utilizes two core strategies for enrichment: (1) topic-based enrichment using extracted named entities and (2) feature-based enrichment where comments are automatically characterized by a set of semantic facets. “*Facet Extraction and Ranking*”, the Facet Extraction and Ranking component can operate on semantically rich comments to cluster comments adaptively along multiple explicit semantic facets (such as subjective comments, informative comments, or comments related to a specific Topic, etc). This component enables an individual user to explore facets, select a combination of facets, and rank comments with regard to an individual user’s preferences. “*Feedback Collector and Optimization*”, the goal of this component is to enable users to provide implicit and explicit feedback. This feedback enables active learning which allows: (1) the ordering of facets and extraction of comments in accordance with user’s interests, and (2) improving the clustering and ranking of comments. “*Baseline Usefulness Model*”, the baseline component of the framework is the “usefulness” classifier which predicts whether each unlabeled enriched comment is useful or non-useful. The framework uses this model as the baseline if the user does not explicitly or implicitly give the system feedback.

The development of a Web-based user interface implementation of the framework allowed us to evaluate different faceted moderation strategies and the proposed framework. We found that adaptive faceted moderation performs better compared to the most commonly used default ranking method and allows users to find interesting comments based on their preferences. Based on our experimental results, we conclude that adaptive faceted ranking performs significantly better than reverse-chronological ranking strategies. There are substantial benefits which include clustering each element of a comment along multiple explicit semantic facets rather than in a single topic order and extracting more objective facets rather than subjective or topic facets.

We will explore the personalized ordering of facets and the personalized faceted ranking strategies by using the active learning in future work. These adapt the faceted ranking to a given user profile that is generated by the user modeling, which may facilitate personalized ranking and improve the facets selection.

Chapter 8

Appendices

8.1 Appendix1 – Experimental Datasets of Related Work

Table 8.1 provides a short overview of main contributions and experimental datasets of each related work, discussed in Chapter 2. In the table, ‘C’ indicates Community-based, ‘S’ indicates Single-user, and ‘CS’ indicates a case study. For approaches related to the single-user, the third column of Table 8.1, instead of the value, shows the related proposed method.

Table 8.1: Short overview of main contributions and experimental datasets of each related work. For framework, ‘C’ indicates Community-based, ‘S’ indicates Single-user, and ‘CS’ indicates a case study

Values	Sys	References	Experimental Dataset
Postings in Micro-blogging			
Credibility	C	“Information credibility on twitter” (WWW 2011) [Castillo et al., 2011]	collected a set of messages related to news events (10,000 tweets) from Twitter and used the Mechanical Turk coders for labeling credibility of tweets.

Credibility	C	“Tweeting is believing?: understanding microblog credibility perceptions” (CHI 2012) [Morris et al., 2012]	conducted a survey with selected participants.
Relevance	C	“Selecting Quality Twitter Content for Events” (ICWSM 2011) [Becker et al., 2011b], “Identifying Content for Planned Events Across Social Media Sites” (WSDM 2012) [Becker et al., 2012]	compiled a dataset of events utilizing content posted between 13th May, 2011 and June 11, 2011 on four different platforms for aggregating events: “Last.fm events”, “EventBrite”, “LinkedIn events”, and “Facebook events”. Furthermore, gathered social media posts for the events from three social media platforms: “Twitter”, “YouTube”, and “Flickr”.
Popularity	C	“Predicting popular messages in Twitter” (WWW 2011) [Hong et al., 2011]	collected messages in November and December 2009 and social contexts of the users which were active at that time. The dataset contains 10,612,601 messages and 2,541,178 users and popularity is calculated by the number of retweets.
Quality	C	“Making sense of twitter” (ISWC 2010) [Laniado and Mika, 2010]	collected a dataset from Twitter during the month of November 2009, which contains 539,432,680 messages.
Relevance	C	“What Makes a Tweet Relevant for a Topic?” (WWW 2012) [Tao et al., 2012]	used the Twitter corpus which had been used in the microblog track of TREC 2011.
Relevance	C	“Beyond trending topics: Real-world event identification on Twitter” (ICWSM 2011) [Becker et al., 2011a]	used a dataset from Twitter, consisting about 2,600,000 messages from February 2010 and used human coders to label clusters for both the training and testing phases of the experiments.
Attention	C	“Predicting Discussions on the Social Semantic Web” (ESWC 2011) [Rowe et al., 2011]	used two online datasets of tweets (http://infochimps.com/datasets/)
Interactive	S	“Finding and assessing social media information sources in the context of journalism” (CHI 2012) [Diakopoulos et al., 2012], “Unfolding the event landscape on twitter: classification and exploration of user categories” (CSCW 2012) [De Choudhury et al., 2012]	collected 3 sets of tweeter posts related to events: (1) a local meeting (similar to a conference) in New York City in July 9th, 2011 which contains 67 sources and 277 Twitter posts. (2) “Tottenham riots” in England on August 7th, 2011 which contains 402 sources and 551 posts. (3) “Tottenham and Joplin” event on May 22nd, 2011 which contains 7,263 sources and 12,595 posts.
Personalized	S	“Leveraging Noisy Lists for Social Feed Ranking” (ICWSM 2013) [Burgess et al., 2013]	first, (1) manually collected a set of 10 lists from www.listorius.com — these cover different topics such as “computer science”, “cooking”, etc. Each list includes 300 and 500 users. These users built a set of seed users who were perhaps members of many lists, (2) acquired all lists that contained any of the seed users, (3) found the creator of each list and provided a set of nearly 400,000 users, and (4) randomly sampled 100 users from the follower set.

Personalized	S	“User oriented tweet ranking: a filtering approach to microblogs” (CIKM 2011) [Uysal and Croft, 2011]	(1) crawled 242 ordinary seed users, all their followers and tweets, (2) for each seed user, randomly selected 100 tweets that would appear on her Twitter feed. In total 24,200 tweets, 2,547 of which were retweeted by the seed users.
Personalized, Adaptive	S	“Whoo.ly: facilitating information seeking for hyperlocal communities using social media” (CHI 2013) [Hu et al., 2013]	using a within-subjects comparison of Whoo.ly and Twitter where users completed tasks to search for information on each platform and then provided feedback.
Personalized	S	“Predicting the Importance of Newsfeed Posts and Social Network Friends” (AAAI 2010) [Paek et al., 2010]	conducted a laboratory study with selected Facebook users (24 users). Participants were asked to rate the importance of their newsfeed posts and friends.
Adaptive	S	“Leveraging the Semantics of Tweets for Adaptive Faceted Search on Twitter” (ISWC 2011) [Abel et al., 2011]	collected a set of tweets by monitoring the Twitter activities of more than 20,000 Twitter (starting from popular Twitter accounts in the news domain and then extended the set of accounts with users who replied or re-tweeted messages) for four months starting from November 2010 (in total more than 30 million Twitter messages were collected).
Interactive	S	“Eddi: interactive topic-based browsing of social status streams” (UIST 2010) [Bernstein et al., 2010]	conducted a laboratory study for evaluating to what extent the Eddi performs better for browsing personal feed than standard reverse-chronological ranking strategy.
Product Review			
Spam	C	“Analyzing and Detecting Review Spam” (ICDM 2007) [Jindal and Liu,], “Opinion spam and analysis” (WSDM 2008) [Jindal and Liu, 2008]	crawled product reviews from amazon.com, including 5.8 million reviews written on 6.7 million products by 214 reviewers.
Deceptive	C	“Estimating the prevalence of deception in online review communities” (WWW 2012) [Ott et al., 2012], “Finding deceptive opinion spam by any stretch of the imagination” (ACL 2011) [Ott et al., 2011]	created a balanced set of 800 training reviews, containing 400 truthful reviews from six online review communities, and 400 gold-standard deceptive reviews from trained Amazon Mechanical Turk coders.
Deceptive	C	“Comparison of Deceptive and Truthful Travel Reviews” (ENTER 2009) [Yoo and Gretzel, 2009]	crawled 40 deceptive hotel reviews from students who studied tourism marketing and extracted truthful reviews from the TripAdvisor.com.

Helpfulness	C	“Designing novel review ranking systems: predicting the usefulness and impact of reviews” (EC 2007) [Ghose and Ipeirotis, 2007], “Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics” (TKDE 2011) [Ghose and Ipeirotis, 2011]	create a dataset of product reviews and related information about prices of product prices and sales and sales rankings from Amazon.com.
Helpfulness	C	“Automatically assessing review helpfulness” (EMNLP 2006) [Kim et al., 2006a]	collected product reviews related to two product categories: “MP3 Players” and “Digital Cameras” from Amazon.com.
Helpfulness	C	“Low-Quality Product Review Detection in Opinion Summarization” (EMNLP 2007) [Liu et al., 2007]	built a ground-truth from the Amazon data set. Collected 4,909 reviews and then hired two human coders to label the reviews.
Helpfulness	C	“Exploiting social context for review quality prediction” (WWW 2012) [Lu et al., 2010]	collected reviews, reviewers, and ratings until May 2009 for all products in three groups: “Cell-phones”, “Beauty”, and “Digital Cameras” from Ciao UK. For measuring a value of review quality (as gold standard), average rating of the reviews is used (a real value between 0 and 5).
Helpfulness	C	“Learning to recommend helpful hotel reviews” (RecSys 2009) [O’Mahony and Smyth, 2009]	built two datasets by crawling all reviews before April 2009 from TripAdvisor. Reviews were selected from users who had reviewed at least one hotel in “Chicago” or “Las Vegas” and had received a minimum of five (either positive or negative) opinion votes.
Helpfulness	C	“RevRank: A Fully Unsupervised Algorithm for Selecting the Most Helpful Book Reviews” (ICWSM 2009) [Tsur and Rappoport, 2009]	tested their system on reviews written for five books with five different genres from Amazon.com. Labeled each review by three different human coders.
Helpfulness	C	“Utility scoring of product reviews” (CIKM 2006) [Zhang and Varadarajan, 2006]	used Amazon.com to obtain a set of reviews.
Helpfulness	CS	“How opinions are received by online communities: a case study on amazon.com helpfulness votes” (WWW 2009) [Danescu-Niculescu-Mizil et al., 2009]	compiled a dataset which contained 4 million reviews (which received at least 10 helpfulness votes) on 675,000 books from Amazon.com.
Comments on Media Objects and Online Forums			
Usefulness	C	“How useful are your comments?: analyzing and predicting youtube comments and comment ratings” (WWW 2010) [Siersdorfer et al., 2010]	created a test collection by obtaining 756 keywords, searched for “related videos”, and gathered the first 500 comments for the video, along with their authors, timestamps and comment ratings for each video.

Usefulness	C	“Properties, Prediction, and Prevalence of Useful User-generated Comments for Descriptive Annotation of Social Media Objects” (ICWSM 2013) [Momeni et al., 2013a], “Identification of Useful User Comments in Social Media: A Case Study on Flickr Commons” (JCDL 2013) [Momeni et al., 2013b]	searched YouTube for videos and Flickr for photos related to three types of topics: “event”, “person”, and “place”. Topics were extracted from the history timeline of the 20th century provided by About.com. 91,778 comments from YouTube and 33,273 comments from Flickr were crawled and used CrowdFlower coders for labeling useful comments
Quality	C	“Ranking Comments on the Social Web” (CSE 2009) [Hsu et al., 2009]	compiled a corpus by crawling stories of the previous 365 days in November 2008 from Digg which contained 9,000 Digg stories and 247,004 comments posted by 47,084 individual users.
Attention	C	“What Catches Your Attention? An Empirical Study of Attention Patterns in Community Forums” (ICWSM2012) [Wagner et al., 2012b], “Ignorance isn’t Bliss: An Empirical Analysis of Attention Patterns in Online Communities” (SocialCom 2012) [Wagner et al., 2012a]	used all data published in the year 2006 from Boards which contained 10 dataset from 10 different community forums.
Credibility	C	“Finding Credible Information Sources in Social Networks Based on Content and Social Structure” (SocialCom 2011) [Canini et al., 2011]	selected five various domains of expertise and then selected by hand 10 Twitter users with high relevancy and expertise for those domains. For selecting relevant users a Twitter service, WeFollow were used.
Quality	C	“Automatically assessing the post quality in online discussions on software” (ACL 2007) [Weimer et al., 2007]	compiled a dataset by collecting posts on the “Software” category of Nabble.com, which contained 1968 rated posts in 1788 threads from 497 forums.
Quality	C	“Slash(dot) and burn: distributed moderation in a large online conversation space” (CHI 2004) [Lampe and Resnick, 2004]	created a dataset from usage logs of slashdot.org between May 31, 2003 to July 30, 2003. These logs contained the “karma” scores of users, and status of users (regular or paid users). The dataset contained 489,948 comments, 293,608 moderations, and 1,576,937 meta-moderations.
Popularity	C	“Predicting the popularity of online content” (ACM COMM 2010) [Szabo and Huberman, 2010]	assembled a dataset which contained 29 million Digg stories written by 560,000 users on 2.7 million posts. Also gathered “view-count time series” on 7,146 selected YouTube videos.
Quality	C	“Automatic Moderation of Comments in a Large Online Journalistic Environment” (ICWSM 2007) [Velooso et al., 2007]	collected 301,278 comments on 472 stories, which were published on Slashdot

Request	CS	“Introductions and Requests: Rhetorical Strategies That Elicit Response in Online Communities Moira” (C&T2007) [Burke et al., 2007]	conduct a series of studies related to the impact of two rhetorical strategies on community responsiveness: “Introduction” and “Request”
Personalized	CS	“Towards quality discourse in on-line news comments” (CSCW2011) [Diakopoulos and Naaman, 2011]	conducted interviewes with 18 people (including editors, reporters, and moderators).
Personalized	S	“Towards quality discourse in on-line news comments” (CSCW 2011) [Lampe et al., 2007]	assembled a dataset from slashdot logs, which contained factors that affected how comments were displayed (such as viewing preferences, etc), a general user information (such as user history and reputation level), and information related to requests of a user
Personalized	S	“Learning to rank social update streams” (SIGIR 2012)[Hong et al., 2012]	created a dataset from the structural data and posts on 99 groups from June 2003 to February 2005 from Usenet.
Questions and Answers in QAC			
Credibility	C	“Learning to recognize reliable users and content in social media with coupled mutual reinforcement” (WWW 2009) [Bian et al., 2009]	used the TREC Q&A queries, searched for these on Yahoo! Answers and crawled questions, answers, and related user information.
Quality	C	“Finding high-quality content in social media with an application to community-based question answering” (WSDM 2008) [Agichtein et al., 2008]	created a dataset containing 8,366 Q&A pairs and 6,665 questions from Yahoo! Answers. Acquired basic usage features from a question thread (page views or clicks) .
Objectivity	C	“CoCQA: co-training over questions and answers with an application to predicting question subjectivity orientation” (EMNLP 2008) [Li et al., 2008]	created a dataset with 1,000 questions from Yahoo! Answers by crawling more than 30,000 questions from top-level categories and randomly selecting 200 questions from each category. Finally gathered labeled for questions using the AmazonÖs Mechanical Turk coders.
Conversational	CS	“Facts or friends?: distinguishing informational and conversational questions in social Q&A sites” (CHI 2009) [Harper et al., 2009]	built a dataset (including full text, names of category, user identifiers, and timestamps) from three Q&A sites (Yahoo, Answerbag, Metafilter) and developed an online coding tool making use of available volunteers for manual coding.
Quality	C	“Predicting information seeker satisfaction in community question answering” (SIGIR 2008) [Liu et al., 2008]	collected a dataset using a snapshot of Yahoo! Answers in 2008, which contained 216,170 questions.
Quality	C	“A framework to predict the quality of answers with non-textual features” (SIGIR 2006) [Jeon et al., 2006]	assembled a dataset by crawling 6.8 million Q&A pairs from the Naver Q&A and randomly chose 894 Q&A pairs from the Naver collection and judged the quality of the answers using human coders.

Relevance	C	“Finding the right facts in the crowd: factoid question answering over social media” (WWW 2008) [Bian et al., 2008]	obtained the 1,250 TREC factoid questions that included at least one similar question from the Yahoo! Answers archive from seven years of the TREC Q&A track evaluations (1999–2006) and labeled the data in two steps: (1) obtaining the TREC factoid answer patterns, (2) independently and manually labeled in order to validate the automatic labels obtained from TREC factoid answer patterns.
Fact	C	“Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences” (EMNLP 2003) [Yu and Hatzivassiloglou, 2003]	used the TREC2 8, 9, and 11 collections which included six different newswire sources — 173,252 articles from Wall Street Journal (WSJ) which included “Editorial”, “Letter to Editor”, “Business”, and “News” from 1987 to 1992 — and randomly picked up 2,000 articles from each category.
Quality	CS	“Predictors of answer quality in online Q&A sites” (CHI 2008) [Harper et al., 2008]	conducted controlled field study of questions and answers from three Q&A sites (Yahoo, Answerbag, Metafilter)
User-Generated Tags			
Quality	C	“Resolving tag ambiguity” (ACM MM 2008) [Weinberger et al., 2008]	collected tags on 102 million Flickr photos which were uploaded between February 2004 and December 2007 and each photo included at least one tag.
Quality	C	“The quest for quality tags” (GROUP 2007) [Sen et al., 2007]	collected 52,814 tags in 9,055 distinct tag sets from MovieLens3 movie recommendation system.
Quality	CS	“What do you call it?: a comparison of library-created and user-created tags” (JCDL 2011) [Hall and Zarro, 2011]	compared the metadata created by two different communities, the ipl2 digital library, and the social tagging system Delicious.
Quality	CS	“What drives content tagging: the case of photos on Flickr” (CHI 2008) [Nov et al., 2008]	conducted a quantitative study for examining what motivation factors correlated with tagging levels, using Flickr tags.
Quality	CS	“Flickr tag recommendation based on collective knowledge” (WWW 2008) [Sigurbjörnsson and van Zwol, 2008]	used a random set (52 million) from Flickr photos uploaded between February 2004 and June 2007 and each photo had at least one user-defined tag.

8.2 Appendix2 – Online Evaluation Instruction

This section shows the online instruction page that trained the evaluation participants for evaluating the framework.

Some online social media objects (such as YouTube videos or online News articles)

include useful and interesting comments. However, due to the huge number of comments, it is often time-consuming and challenging to identify useful comments.

AMOWA is a Web service, which provides automated support for faceted browsing and ranking of social media comments. The service extracts the main topics discussed in the comments and characterizes each comment according to different facets (e.g., subjectivity, emotional level, Informative, and offensiveness of comments). You can use the system to explore the comments through a combination of topics and facets that will allow you to filter and extract those that fit your interests.

The goal of this study is the evaluation of the AMOWA service. In order to evaluate the service you should (1) use the AMOWA service and rank comments for a video based on time (reverse-chronological order) (2) use the AMOWA service and rank comments for the same video accordance with your preferences by selecting provided facets and topics. (3) for each ranking strategy (reverse-chronological or faceted ranking), vote for each comment if the comment is:

- 1. Interesting: If it contains interesting content for you personally and not necessarily for others users.*
- 2. Relevant: If it contains relevant content to your selected facets and topics. Please note that the comment does not necessarily have to be directly relevant to the video content. The facets and topics refer to the comments, which not necessarily match the topic of the video.*

Please follow the instruction below carefully:

PLEASE NOTE: IN ORDER TO GET APPROVAL FOR THIS TASK YOU SHOULD COMPLETE ALL 6 STEPS AND FOUR SUB-STEPS (4a – 4d).

- 1. Create an account using the following address:
<http://amowa.cs.univie.ac.at:8080/Frontend/register.html>. Please note it is important that you use a real e-mail address as user name, so we can reach you for payment.*

2. *Use the AMOWA service:*

<http://amowa.cs.univie.ac.at:8080/Frontend/> Start by ranking comments for the selected video based on time (YouTube default). To do this you should:

- *(a) select the title of a video from the Option Box and click the load-button. Some videos have lots of comments, therefore the loading will take up to 40 seconds, please DO NOT press the loading button again.*
- *(b) the system will show the video along with the list of topics and facets related to the comments of the video on the left.*
- *(c) watch the selected video.*
- *(d) For reverse-chronological order, select “reverse-chronological” and you will immediately see comments on the right side.*

3. *For each comment listed on the right side, you will see two voting choices (“interesting” and “relevant”). For the reverse-chronological ranking, just vote on all or at least the first 30 comments, if the comment is interesting for you or not. You do not need to vote on “relevant” for the reverse-chronological order. After voting for at least 30 comments, click on “Save votes” at the end of the list. Please note that you should vote for at least 30 comments, otherwise the system does not let you save your votes.*

4. *You have just completed the “reverse-chronological” ranking step. Now, use the AMOWA service to rank comments based on your preferences by selecting different topic(s), facet(s), or both. Please follow the four steps below:*

- *(a) Select one or more Topics in accordance with your preferences and vote on all or at least the first 30 comments in the list : “is the comment interesting for you or not” and also vote “is the comment relevant to your topic selection”. After finish voting, click on Save votes at the end of the list.*
- *(b) Select one or more Facets from these Facets (Religious referenced, Subjective opinion, Affective, Offensive, Anger oriented, Sad oriented) in accordance with your preferences and vote on all or at least the first 30*

comments in the list: “is the comment interesting for you or not” and also vote on “is the comment relevant to your selection of facets”. After finish voting, click on Save votes at the end of the list.

- *(c) Then select one or more Facets from these Facets (Informative, Video timestamp) in accordance with your preferences and vote on all or at least the first 30 comments in the list : “is the comment interesting for you or not” and also vote on “is the comment relevant to your selection of facets”. After finish voting, click on Save votes at the end of the list.*
- *(d) Select a combination of a Facet(s) and/or a Topic(s) in accordance with your preferences and vote on all or at least the first 30 comments in the list: “is the comment interesting for you or not” and “is the comment relevant to your facets and/or topics selection”. After finish voting, click on Save votes at the end of the list.*

5. *Having completed the 4 steps above, repeat the steps 2 to 4 for the second video.*
6. *After completing all ranking steps for both videos please use the following online form and provide us your background and feedback*
<http://amowa.cs.univie.ac.at:8080/Frontend/feedback.html>. Please note this is a very important step, which helps us check and approve your contributions and we need this information to be able to send you the gift certificate.

Bibliography

- [Abel et al., 2011] Abel, F., Celik, I., Houben, G.-J., and Siehndel, P. (2011). Leveraging the semantics of tweets for adaptive faceted search on twitter. In *Proceedings of the 10th International Conference on The Semantic Web - Volume Part I*, ISWC’11, pages 1–17, Berlin, Heidelberg. Springer-Verlag.
- [Agichtein et al., 2008] Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G., Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. (2008). Finding high-quality content in social media with an application to community-based question answering. In *Proceedings of the Second ACM international conference on Web search and data mining*.
- [Ames and Naaman, 2007] Ames, M. and Naaman, M. (2007). Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’07, New York, NY, USA. ACM.
- [Asselin et al., 2011] Asselin, M., Dobson, T., Meyers, E. M., Teixeira, C., and Ham, L. (2011). Learning from youtube: An analysis of information literacy in user discourse. In *Proceedings of the 2011 iConference*, iConference ’11, pages 640–642, New York, NY, USA. ACM.
- [Baccianella and Sebastiani, 2010] Baccianella, A. E. S. and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC’10)*, Valletta, Malta. European Language Resources Association (ELRA).

- [Becker et al., 2012] Becker, H., Iter, D., Naaman, M., and Gravano, L. (2012). 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.
- [Becker et al., 2011a] Becker, H., Naaman, M., and Gravano, L. (2011a). Beyond trending topics: Real-world event identification on twitter. In *Fifth International AAAI Conference on Weblogs and Social Media*.
- [Becker et al., 2011b] Becker, H., Naaman, M., and Gravano, L. (2011b). Selecting quality twitter content for events. In Adamic, L. A., Baeza-Yates, R. A., and Counts, S., editors, *Fifth International AAAI Conference on Weblogs and Social Media*. The AAAI Press.
- [Beenen et al., 2004] Beenen, G., Ling, K., Wang, X., Chang, K., Frankowski, D., Resnick, P., and Kraut, R. E. (2004). Using social psychology to motivate contributions to online communities. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, CSCW '04, pages 212–221, New York, NY, USA. ACM.
- [Bernstein et al., 2010] Bernstein, M. S., Suh, B., Hong, L., Chen, J., Kairam, S., and Chi, E. H. (2010). Eddi: interactive topic-based browsing of social status streams. In *Proceedings of the 23rd annual ACM symposium on User interface software and technology*, UIST '10, pages 303–312, New York, NY, USA. ACM.
- [Bian et al., 2008] Bian, J., Liu, Y., Agichtein, E., and Zha, H. (2008). Finding the right facts in the crowd: factoid question answering over social media. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 467–476, New York, NY, USA. ACM.
- [Bian et al., 2009] Bian, J., Liu, Y., Zhou, D., Agichtein, E., and Zha, H. (2009). Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 51–60, New York, NY, USA. ACM.

- [Blei et al., 2003] Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *the Journal of machine Learning res.*
- [Blum and Mitchell, 1998] Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory, COLT' 98*, pages 92–100, New York, NY, USA. ACM.
- [Boyd-Graber et al., 2009] Boyd-Graber, J., Chang, J., Gerrish, S., Wang, C., and Blei, D. (2009). Reading tea leaves: How humans interpret topic models. In *Proceedings of the Neural Information Processing Systems (NIPS 2010)*.
- [Burgess et al., 2013] Burgess, M., Mazzia, A., Adar, E., and Cafarella, M. (2013). Leveraging noisy lists for social feed ranking. In *The 7TH International AAAI Conference On Weblogs And Social Media (ICWSM2013)*, Boston, USA. AAAI.
- [Burke et al., 2007] Burke, M., Joyce, E., Kim, T., An, V., and Kraut, R. (2007). Introductions and requests: Rhetorical strategies that elicit response in online communities. In *C&T '07: Third International Conference on Communities & Technologies 2007, East*, pages 21–40.
- [Canini et al., 2011] Canini, K. R., Suh, B., and Pirolli, P. L. (2011). Finding Credible Information Sources in Social Networks Based on Content and Social Structure. In *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing*, pages 1–8. IEEE.
- [Castillo et al., 2011] Castillo, C., Mendoza, M., and Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web, WWW '11*.
- [Chen et al., 2010] Chen, J., Nairn, R., Nelson, L., Bernstein, M., and Chi, E. (2010). Short and tweet: Experiments on recommending content from information streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '10*, pages 1185–1194, New York, NY, USA. ACM.

- [Choudhury et al., 2012] Choudhury, M. D., Counts, S., and Gamon, M. (2012). Not all moods are created equal! exploring human emotional states in social media. In *In Sixth International AAAI Conference on Weblogs and Social Media*.
- [Danescu-Niculescu-Mizil et al., 2009] Danescu-Niculescu-Mizil, C., Kossinets, G., Kleinberg, J., and Lee, L. (2009). How opinions are received by online communities: a case study on amazon.com helpfulness votes. In *Proceedings of the 18th international conference on World wide web, WWW '09*.
- [De Choudhury et al., 2012] De Choudhury, M., Diakopoulos, N., and Naaman, M. (2012). Unfolding the event landscape on twitter: classification and exploration of user categories. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, CSCW '12*, pages 241–244, New York, NY, USA. ACM.
- [Diakopoulos et al., 2012] Diakopoulos, N., De Choudhury, M., and Naaman, M. (2012). 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.
- [Diakopoulos and Naaman, 2011] Diakopoulos, N. and Naaman, M. (2011). Towards quality discourse in online news comments. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work, CSCW '11*, pages 133–142, New York, NY, USA. ACM.
- [Ghose and Ipeirotis, 2007] Ghose, A. and Ipeirotis, P. G. (2007). Designing novel review ranking systems: predicting the usefulness and impact of reviews. In *ICEC '07: Proceedings of the ninth international conference on Electronic commerce*.
- [Ghose and Ipeirotis, 2011] Ghose, A. and Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10):1498–1512.

- [Ghosh, 2012] Ghosh, A. (2012). Social computing and user-generated content: A game-theoretic approach. In *ACM SIGecom Exchanges*, SIGecom '12, pages Vol. 12, No. 2, New York, NY, USA. ACM.
- [Ghosh and Hummel, 2011] Ghosh, A. and Hummel, P. (2011). A game-theoretic analysis of rank-order mechanisms for user-generated content. In *Proceedings of the 12th ACM conference on Electronic commerce*, EC '11, pages 189–198, New York, NY, USA. ACM.
- [Ghosh and McAfee, 2011] Ghosh, A. and McAfee, P. (2011). Incentivizing high-quality user-generated content. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 137–146, New York, NY, USA. ACM.
- [Ghosh and McAfee, 2012] Ghosh, A. and McAfee, P. (2012). Crowdsourcing with endogenous entry. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 999–1008, New York, NY, USA. ACM.
- [Golder and Huberman, 2006] Golder, S. A. and Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208.
- [Gunning, 1952] Gunning, R. (1952). *The Technique of Clear Writing*. McGraw-Hill, New York.
- [Hall and Zarro, 2011] Hall, C. E. and Zarro, M. A. (2011). 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, pages 53–56, New York, NY, USA. ACM.
- [Halpin et al., 2007] Halpin, H., Robu, V., and Shepherd, H. (2007). The complex dynamics of collaborative tagging. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07.
- [Harper et al., 2009] Harper, F. M., Moy, D., and Konstan, J. A. (2009). 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.

- [Harper et al., 2008] Harper, F. M., Raban, D., Rafaeli, S., and Konstan, J. A. (2008). Predictors of answer quality in online q&a sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 865–874, New York, NY, USA. ACM.
- [Haslhofer et al., 2010] Haslhofer, B., Momeni, E., Gay, M., and Simon, R. (2010). Augmenting europeana content with linked data resources. In *Linked Data Triplification Challenge, co-located with I-Semantics 2010*, Proceedings of the 6th International Conference on Semantic Systems, pages 40:1–40:3, New York, NY, USA. ACM.
- [Haslhofer et al., 2013] Haslhofer, B., Robitza, W., Guimbretiere, F., and Lagoze, C. (2013). Semantic tagging on historical maps. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, pages 148–157, New York, NY, USA. ACM.
- [Hearst, 2009] Hearst, M. A. (2009). *Search User Interfaces*. Cambridge University Press, New York, NY, USA, 1st edition.
- [Hevner et al., 2004] Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004). Design science in information systems research. *Journal of MIS Quarterly*, 28(1):75–105.
- [Hong et al., 2012] Hong, L., Bekkerman, R., Adler, J., and Davison, B. D. (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, pages 651–660, New York, NY, USA. ACM.
- [Hong et al., 2010] Hong, L., Convertino, G., Suh, B., Chi, E. H., and Kairam, S. (2010). Feedwinnow: Layering structures over collections of information streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 947–950, New York, NY, USA. ACM.
- [Hong et al., 2011] Hong, L., Dan, O., and Davison, B. D. (2011). Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, pages 57–58, New York, NY, USA. ACM.

- [Hsu et al., 2009] Hsu, C.-F., Khabiri, E., and Caverlee, J. (2009). Ranking comments on the social web. In *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04*, CSE '09, pages 90–97, Washington, DC, USA. IEEE Computer Society.
- [Hu et al., 2013] Hu, Y., Farnham, S. D., and Monroy-Hernández, A. (2013). Whoo.ly: facilitating information seeking for hyperlocal communities using social media. In *CHI*, pages 3481–3490.
- [Huberman et al., 2009] Huberman, B. A., Romero, D. M., and Wu, F. (2009). Crowdsourcing, attention and productivity. *J. Inf. Sci.*, 35(6):758–765.
- [Jeon et al., 2006] Jeon, J., Croft, W. B., Lee, J. H., and Park, S. (2006). A framework to predict the quality of answers with non-textual features. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 228–235, New York, NY, USA. ACM.
- [Jindal and Liu,] Jindal, N. and Liu, B. Analyzing and detecting review spam. In *ICDM '07: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*.
- [Jindal and Liu, 2008] Jindal, N. and Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 219–230, New York, NY, USA. ACM.
- [Kennedy et al., 2007] Kennedy, L., Naaman, M., Ahern, S., Nair, R., and Rattenbury, T. (2007). How flickr helps us make sense of the world: context and content in community-contributed media collections. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, pages 631–640, New York, NY, USA. ACM.
- [Kim et al., 2006a] Kim, S.-M., Pantel, P., Chklovski, T., and Pennacchiotti, M. (2006a). Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06.

- [Kim et al., 2006b] Kim, S.-M., Pantel, P., Chklovski, T., and Pennacchiotti, M. (2006b). Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06.
- [Kitchenham, 2004] Kitchenham, B. (2004). Procedures for performing systematic reviews. Technical report, Department of Computer Science, Keele University.
- [Koren et al., 2008] Koren, J., Zhang, Y., and Liu, X. (2008). Personalized interactive faceted search. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 477–486, New York, NY, USA. ACM.
- [Lampe and Resnick, 2004] Lampe, C. and Resnick, P. (2004). Slash(dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 543–550, New York, NY, USA. ACM.
- [Lampe et al., 2007] Lampe, C. A., Johnston, E., and Resnick, P. (2007). Follow the reader: filtering comments on slashdot. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 1253–1262, New York, NY, USA. ACM.
- [Laniado and Mika, 2010] Laniado, D. and Mika, P. (2010). Making sense of twitter. In *Proceedings of the 9th international semantic web conference on The semantic web - Volume Part I*, ISWC'10, pages 470–485, Berlin, Heidelberg. Springer-Verlag.
- [Li et al., 2008] Li, B., Liu, Y., and Agichtein, E. (2008). Cocqa: co-training over questions and answers with an application to predicting question subjectivity orientation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 937–946, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Lieberman and Lempel, 2012] Lieberman, S. and Lempel, R. (2012). Approximately optimal facet selection. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, SAC '12, pages 702–708, New York, NY, USA. ACM.

- [Liu et al., 2007] Liu, J., Cao, Y., Lin, C. Y., Huang, Y., and Zhou, M. (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)*.
- [Liu et al., 2008] Liu, Y., Bian, J., and Agichtein, E. (2008). 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.
- [Lu et al., 2010] Lu, Y., Tsaparas, P., Ntoulas, A., and Polanyi, L. (2010). Exploiting social context for review quality prediction. In *Proceedings of the 19th international conference on World wide web*, WWW '10.
- [Momeni, 2010] Momeni, E. (2010). Towards (semi-)automatic moderation of social web annotations. In *Proceedings of the Second IEEE International Conference on Social Computing, 2010 (SocialCom2010)*. IEEE.
- [Momeni, 2012] Momeni, E. (2012). Semi-automatic semantic moderation of web annotations. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '12 Companion, pages 167–172, New York, NY, USA. ACM.
- [Momeni and Cardie, 2014] Momeni, E. and Cardie, C. (2014). A survey on assessment and ranking methodologies for user-generated content on web. under review.
- [Momeni et al., 2013a] Momeni, E., Cardie, C., and Ott, M. (2013a). 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)*, Boston, USA. AAAI.
- [Momeni et al., 2014a] Momeni, E., Hauff, C., Braendle, S., and Adar, E. (2014a). Multi-faceted adaptive ranking of social media comments. under review.

- [Momeni and Sageder, 2013] Momeni, E. and Sageder, G. (2013). An empirical analysis of characteristics of useful comments in social media. In *Proceedings of the ACM Web Science*, WebSci2013.
- [Momeni et al., 2013b] Momeni, E., Tao, K., Haslhofer, B., and Houben, G.-J. (2013b). Identification of useful user comments in social media: A case study on flickr commons. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '13, pages 1–10, New York, NY, USA. ACM.
- [Momeni et al., 2014b] Momeni, E., Tao, K., Haslhofer, B., and Houben, G.-J. (2014b). Sifting useful comments from flickr commons and youtube. In *International Journal on Digital Libraries*, IJDL '14, New York, NY, USA. Springer.
- [Morris et al., 2012] Morris, M. R., Counts, S., Roseway, A., Hoff, A., and Schwarz, J. (2012). Tweeting is believing?: understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, pages 441–450, New York, NY, USA. ACM.
- [Nam et al., 2009] Nam, K. K., Ackerman, M. S., and Adamic, L. A. (2009). Questions in, knowledge in?: a study of naver's question answering community. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 779–788, New York, NY, USA. ACM.
- [Nov et al., 2008] Nov, O., Naaman, M., and Ye, C. (2008). What drives content tagging: the case of photos on flickr. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1097–1100, New York, NY, USA. ACM.
- [O'Mahony and Smyth, 2009] O'Mahony, M. P. and Smyth, B. (2009). Learning to recommend helpful hotel reviews. In *Proceedings of the third ACM conference on Recommender systems*, RecSys '09, pages 305–308, New York, NY, USA. ACM.
- [Ott et al., 2012] Ott, M., Cardie, C., and Hancock, J. (2012). Estimating the prevalence of deception in online review communities. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 201–210, New York, NY, USA. ACM.

- [Ott et al., 2011] Ott, M., Choi, Y., Cardie, C., and Hancock, J. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ACL '11, pages 309–319, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Paek et al., 2010] Paek, T., Gamon, M., Counts, S., Chickering, D. M., and Dhesi, A. (2010). Predicting the importance of newsfeed posts and social network friends. In Fox, M. and Poole, D., editors, *AAAI*. AAAI Press.
- [Radev, 2004] Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:2004.
- [Ramage et al., 2010] Ramage, D., Dumais, S. T., and Liebling, D. J. (2010). Characterizing microblogs with topic models. In Cohen, W. W. and Gosling, S., editors, *The 4th International AAAI Conference on Weblog and Social Media*. The AAAI Press.
- [Rangwala and Jamali, 2010] Rangwala, H. and Jamali, S. (2010). Defining a coparticipation network using comments on digg. *IEEE Intelligent Systems*, 25(4):36–45.
- [Rotman et al., 2009] Rotman, D., Golbeck, J., and Preece, J. (2009). The community is where the rapport is – on sense and structure in the youtube community. In *Proceedings of the Fourth International Conference on Communities and Technologies*, C&T '09, pages 41–50, New York, NY, USA. ACM.
- [Rowe et al., 2011] Rowe, M., Angeletou, S., and Alani, H. (2011). Predicting discussions on the social semantic web. In *Extended Semantic Web Conference*, Heraklion, Crete.
- [Sen et al., 2007] Sen, S., Harper, F. M., LaPitz, A., and Riedl, J. (2007). The quest for quality tags. In *Proceedings of the 2007 international ACM conference on Supporting group work*, GROUP '07, pages 361–370, New York, NY, USA. ACM.

- [Shamma et al., 2007] Shamma, D. A., Shaw, R., Shafon, P. L., and Liu, Y. (2007). Watch what i watch: Using community activity to understand content. In *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval*, MIR '07, pages 275–284, New York, NY, USA. ACM.
- [Siersdorfer et al., 2010] Siersdorfer, S., Chelaru, S., Nejdl, W., and San Pedro, J. (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.
- [Sigurbjörnsson and van Zwol, 2008] Sigurbjörnsson, B. and van Zwol, R. (2008). Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 327–336, New York, NY, USA. ACM.
- [Sriram et al., 2010] Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M. (2010). Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 841–842, New York, NY, USA. ACM.
- [Szabo and Huberman, 2010] Szabo, G. and Huberman, B. A. (2010). Predicting the popularity of online content. *Commun. ACM*, 53(8):80–88.
- [Tao et al., 2012] Tao, K., Abel, F., Hauff, C., and Houben, G.-J. (2012). What makes a tweet relevant for a topic? In *Making Sense of Microposts (#MSM2012)*, pages 49–56.
- [Tausczik and Pennebaker, 2010] Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods.
- [Tseng et al., 2012] Tseng, C.-Y., Chen, Y.-J., and Chen, M.-S. (2012). Socfeed-viewer: A novel visualization technique for social news feeds summarization on social network services. In Goble, C. A., Chen, P. P., and Zhang, J., editors, *ICWS*, pages 616–617. IEEE.

- [Tsur and Rappoport, 2009] Tsur, O. and Rappoport, A. (2009). RevRank: A Fully Unsupervised Algorithm for Selecting the Most Helpful Book Reviews. In *International Conference on Weblogs and Social Media (ICWSM09)*.
- [Uysal and Croft, 2011] Uysal, I. and Croft, W. B. (2011). User oriented tweet ranking: a filtering approach to microblogs. In Macdonald, C., Ounis, I., and Ruthven, I., editors, *CIKM*, pages 2261–2264. ACM.
- [Veloso et al., 2007] Veloso, A., Jr., W. M., Macambira, T., Guedes, D., and Almeida, H. (2007). Automatic moderation of comments in a large on-line journalistic environment. In *ICWSM*.
- [Wagner et al., 2012a] Wagner, C., Rowe, M., Strohmaier, M., and Alani, H. (2012a). Ignorance isn’t bliss: An empirical analysis of attention patterns in online communities. In *IEEE International Conference on Social Computing*, Amsterdam, The Netherlands.
- [Wagner et al., 2012b] Wagner, C., Rowe, M., Strohmaier, M., and Alani, H. (2012b). What catches your attention? an empirical study of attention patterns in community forums. In *ICWSM*.
- [Weimer et al., 2007] Weimer, M., Gurevych, I., and Mühlhäuser, M. (2007). Automatically assessing the post quality in online discussions on software. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL ’07, pages 125–128, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Weinberger et al., 2008] Weinberger, K. Q., Slaney, M., and Van Zwol, R. (2008). Resolving tag ambiguity. In *Proceedings of the 16th ACM international conference on Multimedia*, MM ’08, pages 111–120, New York, NY, USA. ACM.
- [Wilson et al., 2005] Wilson, T., Wiebe, J., and Hoffmann, P. (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, pages 347–354, Stroudsburg, PA, USA. Association for Computational Linguistics.

- [Yang et al., 2011] Yang, J., Ackerman, M. S., and Adamic, L. A. (2011). Virtual gifts and guanxi: supporting social exchange in a chinese online community. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work, CSCW '11*, pages 45–54, New York, NY, USA. ACM.
- [Yoo and Gretzel, 2009] Yoo, K. H. and Gretzel, U. (2009). Comparison of deceptive and truthful travel reviews. In *Information and Communication Technologies in Tourism, ENTER 2009, Proceedings of the International Conference in Amsterdam, The Netherlands, 2009*, pages 37–47. Springer.
- [Yu and Hatzivassiloglou, 2003] Yu, H. and Hatzivassiloglou, V. (2003). Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing, EMNLP '03*, pages 129–136, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Zhang and Varadarajan, 2006] Zhang, Z. and Varadarajan, B. (2006). Utility scoring of product reviews. In *Proceedings of the 15th ACM international conference on Information and knowledge management, CIKM '06*, pages 51–57, New York, NY, USA. ACM.

