HELPFULNESS-GUIDED REVIEW SUMMARIZATION

by

Wenting Xiong

B.E. in Information Engineering, Beijing University of Posts
and Telecommunications, 2008

Submitted to the Graduate Faculty of
the Kenneth P. Dietrich School of Arts and Sciences,
Department of Computer Science in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

KENNETH P. DIETRICH SCHOOL OF ARTS AND SCIENCES, DEPARTMENT OF COMPUTER SCIENCE

This dissertation was presented

by

Wenting Xiong

It was defended on

August 5th 2014

and approved by

Diane Litman, Department of Computer Science

Rebecca Hwa, Department of Computer Science

Janyce Wiebe, Department of Computer Science

Jingtao Wang, Department of Computer Science

Christian Schunn, Department of Psychology

Dissertation Director: Diane Litman, Department of Computer Science

HELPFULNESS-GUIDED REVIEW SUMMARIZATION

Wenting Xiong, PhD

University of Pittsburgh, 2014

User-generated online reviews are an important information resource in people's everyday life. As the review volume grows explosively, the ability to automatically identify and summarize useful information from reviews becomes essential in providing analytic services in many review-based applications. While prior work on review summarization focused on different review perspectives (e.g., topics, opinions, sentiment, etc.), the helpfulness of reviews is an important informativeness indicator that has been less frequently explored. In this thesis, we investigate automatic review helpfulness prediction and exploit review helpfulness for review summarization in distinct review domains.

We explore two paths for predicting review helpfulness in a general setting: one is by tailoring existing helpfulness prediction techniques to a new review domain; the other is by using a general representation of review content that reflects review helpfulness across domains. For the first one, we explore educational peer reviews and show how peer-review domain knowledge can be introduced to a helpfulness model developed for product reviews to improve prediction performance. For the second one, we characterize review language usage, content diversity and helpfulness-related topics with respect to different content types using computational linguistic features.

For review summarization, we propose to leverage user-provided helpfulness assessment during content selection in two ways: 1) using the review-level helpfulness ratings directly to filter out unhelpful reviews, 2) developing sentence-level helpfulness features via supervised topic modeling for sentence selection. As a demonstration, we implement our methods based on an extractive multi-document summarization framework and evaluate them in three user studies. Results show that our helpfulness-guided summarizers outperform the baseline in both human and automated evaluation for camera reviews and movie reviews. While for educational peer reviews, the preference for helpfulness depends on student writing performance and prior teaching experience.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

This work would not have been possible without the support of many people, including my parents, my roommates and friends both in China and in the United States. I greatly appreciate the expertise, assistance of my dissertation committee members! And of course, I am sincerely deeply grateful for my dear advisor Diane, who gives me a lot of freedom in the kind of research work that I can do (while divergence sometimes is inevitable), keeps me on track of getting the PhD done, as well as helps me become an independent researcher (if I am now).

While the path of getting a PhD degree could be painful in so many ways, I feel lucky to be able to enjoy the process all the way through, with the help of my friendly lab mates and the support from the whole CS department. It is a really nice experience to spend six years at PITT in Pittsburgh, although it a little bit longer than I expected...

1.0 INTRODUCTION

1.1 BACKGROUND

With the prevalence of Web 2.0 technologies, user generated content such as reviews, blogs, tweets, etc. has received increasing attention in the community of natural language processing (NLP). Areas that are related to social media, data mining and text analytics have harvested many publications in the past decade. Advances in these areas make it possible to develop intelligent information systems for many applications, varying from consolidating online reviews for potential customers, gathering user feedback for commercial companies, providing social question-answering services, or even supporting surveillance and censorship on social media.

Online reviews (e.g., product reviews from Amazon.com, movie reviews from IMDB.com, customer reviews of local services that are directly integrated in searching engines, etc.) are a typical kind of user generated content that serves as an important information resource in people's everyday life. Despite the diversity of things that can be reviewed online, one review usually contains a numeric rating (star rating) and some textual comments. Additionally, a review may have various metadata such as user-provided review helpfulness votes. A lot of topics have been studied on online reviews, including sentiment analysis (Turney, 2002; Pang and Lee, 2004), opinion mining (Hu and Liu, 2004), topic modeling (Titov and McDonald, 2008b), summarization (Carenini et al., 2006), review helpfulness analysis (Kim et al., 2006), and so on. The progress of these topics also interact with each other; techniques proposed in one field also shed light on problems of other fields. For example, topic modeling has been used in fine-grained sentiment analysis (Lu et al., 2011), while sentiment information can be used as supervision for topic inference (Blei and McAuliffe, 2010), and summarization as well (Zhuang et al., 2006; Lerman et al., 2009). As for review helpfulness, although subjectivity and sentiment analysis are found useful for helpfulness prediction (Zeng and Wu, 2013), studies of how to leverage review helpfulness analysis for other tasks are limited.

Among these topics, this thesis focuses on review helpfulness analysis, with special interest in exploiting the helpfulness information for review summarization.

In the past, researchers have investigated what makes an online review perceived helpful in terms of review structure, content (language and semantics), social context (Zeng and Wu, 2013; Mudambi and Schuff, 2010; Tang et al., 2013), etc. The social and economic impact of online consumer reviews (e.g., product reviews) has also been examined. As for automatically predicting review helpfulness, it is often considered as a standard machine learning task, solved in a data driven fashion based on features derived from both review text and review context.

With respect to features derived from the review text, early work found that review length and review unigrams are the most predictive features for product reviews, though using unigrams together with other syntactic features (e.g., statistics of Part-Of-Speech tags) and semantic features (e.g., domain words and sentiment words) decreased the prediction performance. Other studies focused on particular patterns in the textual content to evaluate review helpfulness. Liu et al. (2007) focused on the mentions of product names and evaluation expressions which are mined using opinion mining; Zeng and Wu (2013) also examined product reviews but especially looked at the comparison style, extracted using regular expressions. Tsur and Rappoport (2009) examined reviewers' writing style of book reviews, using syntactic features derived from Part-Of-Speech patterns; Yu et al. (2012) later considered the writing style for movie reviews, though only chose certain POS tags. While such kind of models achieved high performance in their specific machine learning tasks for particular domains, the used features are not directly/indirectly compared in/across domains. Further, these features require sophisticated natural language processing such as opinion mining (Liu et al., 2007; Zeng and Wu, 2013) and parsing (Tsur and Rappoport, 2009; Yu et al., 2012). However, the utility of these sophisticated features is not compared to simple but effective features such as unigrams.

Independently, some semantics and pragmatic features proposed in other NLP fields have also been used for predicting review helpfulness (O'Mahony and Smyth, 2010; Ghose and Ipeirotis, 2011). For example, it has been shown that review readability (e.g., error rate, language formality) is significantly correlated with review helpfulness (O'Mahony and Smyth, 2010). But again, no comprehensive analysis has been reported comparing their effectiveness with low-level features such as review's lexical semantics.

With respect to non-textual features, statistics of review star ratings (sentiment ratings) are shown predictive of product review helpfulness (Kim et al., 2006). In addition, the age of the reviews, the reputation of the reviewer (Liu et al., 2008; Ghose and Ipeirotis, 2011) as well as the interaction between the reviewer and the reader (Lu et al., 2010) are also useful for predicting review helpfulness. In this thesis, we investigate review helpfulness prediction from the perspective of natural language processing, and thus focus on review textual information only.

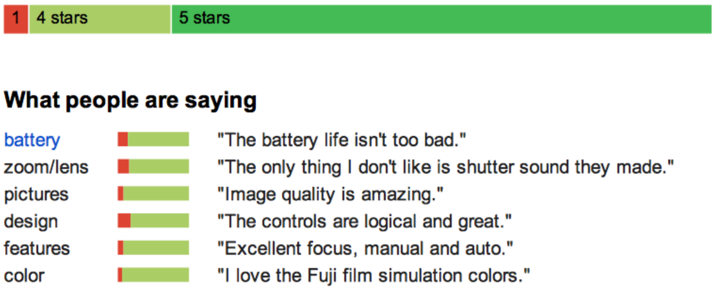
In particular, our research is motivated by the following challenges identified in prior work on review helpfulness introduced above. 1) Existing techniques for review helpfulness prediction are often dedicated to a particular kind of reviews, and the predictive features vary a lot in different review domains. Therefore, it is not clear which model or features to use for new emerging review domains, such as educational peer reviews. 2) Considering the sophistication level of the different feature types, it would be helpful to conduct comprehensive analysis on the effectiveness of high-level content representation versus review unigrams across domains, to justify the need of extra computation. Given the development of related work in other NLP fields, we wonder if review content can be characterized comprehensively, in a way that captures the semantics of review helpfulness in different contexts. If the answer is yes, we can build a general helpfulness model based on this representation to predict review helpfulness in the same way across domains. 3) Furthermore, although related work in subjectivity and sentiment analysis suggests selective use of the content, i.e. identify sentiment lexicons only from the subjective set of documents (or sentences), such kind of content categorization applied before feature engineering has not been explored in the analysis or review helpfulness prediction. 4) In terms of level of analysis granularity, existing studies focus on review helpfulness at the review level, while the helpfulness might be different between sentences within a review.

To address these challenges, this thesis pursues review helpfulness prediction in two different paths. One is through specialization, in which we investigate the feasibility of tailoring existing helpfulness prediction techniques. In particular, we explore a new kind of review – educational peer review, as a case study.

Educational peer reviews are specialized online reviews that have rarely been explored before, but are more and more popular with the development of MOOCs (Massive Open Online Courses). As the size of a class is often quite large on MOOCs, peer assessment and peer review have caught researchers' attention as alternatives to teacher grading on non-multiple choice questions (e.g., writing assignments). To support online peer review activities, web-based peer-review systems have been developed which save instructors a lot of effort in setting up peer-review assignments and managing document assignment. However, there still remains the problem that peer reviews are not always written in a constructive way (Nelson and Schunn, 2009). Therefore, to enhance the effectiveness of existing peer-review activities (and to ultimately improve student learning), computational techniques of assessing peer-review helpfulness are desirable. To specialize a helpfulness model to the education domain, we propose computational linguistic features to capture the educational semantics of helpfulness that has not been attempted before.

The second path that we take to predict review helpfulness is by generalization. Instead of tailoring content extraction to a particular domain, we emphasize generality in describing patterns of review textual content, aiming to model review helpfulness using the same framework in distinct review domains. In this work, we propose a new content representation based on NLP techniques that are shown effective in content analysis of text in other genres. In particular, we examine the effectiveness of our new (textual) content features by comparing it against review unigrams across domains.

In addition to **predicting** review helpfulness, this thesis also explores how to **exploit** the helpfulness information for building review-based applications such as review summarization. Given that the volume of online reviews is growing explosively, the capability of review summarization is also desirable in building review-based information systems such as review-solicit websites and search engines. In the literature, various approaches have been proposed for review summarization, which generally fall into two paradigms.



What people are saying

battery	"The battery life isn't too bad."
zoom/lens	"The only thing I don't like is shutter sound they made."
pictures	"Image quality is amazing."
design	"The controls are logical and great."
features	"Excellent focus, manual and auto."
color	"I love the Fuji film simulation colors."

Figure 1.1: A summary of digital camera reviews generated by Google Shopping.

One paradigm is aspect-based opinion summarization, in which reviews are summarized into a list of a review item's aspects and their corresponding sentiment scores plus a text snippet, such as Figure 1.1. These kinds of table-style summaries are often visualized into charts or graphs to emphasize the summary statistics in an intuitive way (Liu et al., 2005). This approach especially suits building summarization applications on mobile platforms (Huang et al., 2012).

The second paradigm is tailoring standard multi-document summarization methods to the review genre. Summarizers of this kind are able to generate review summaries in natural language that are more like individual reviews, such as the example below:

> *Summary of reviews on Canon G3 (provided by Carenini et al. (2006)):*
>
> Almost all users loved the Canon G3 possibly because some users thought the physical appearance was very good. Furthermore, several users found the manual features and the special features to be very good. Also, some users liked the convenience because some users thought the battery was excellent. Finally, some users found the editing/viewing interface to be good despite the fact that several customers really disliked the viewfinder . However, there were some negative evaluations. Some customers thought the lens was poor even though some customers found the optical zoom capability to be excellent. Most customers thought the quality of the images was very good.

Early work constrained the summarization from the input: applying standard multi-document summarization on only the relevant subset of documents or sentences, such as evaluative text on the same topic (Seki et al., 2006). This requires pre-processing the corpus for a series of NLP tasks including topic classification, content selection and sentiment predictions. As standard multi-document summarization techniques prioritize content with high occurrence frequency, researchers also modify the summarization algorithms for generating different styles of summaries. Some aim to generate a representative sample of opinions, some desire summaries of contradictory opinions, and some are interested in creating contrastive summaries by extracting comparative evaluative arguments. However, these methods tailored their summarization techniques to meet a predefined and limited style, though whether such style matches what users desire can not be always guaranteed.

Review helpfulness metadata and prediction models provide opportunities to get around this issue. Liu et al. (2007) used a helpfulness classifier to detect and filter out unhelpful reviews before the summarization process, though their classifier was trained on expert-provided helpfulness gold-standard. If user provided helpfulness assessment could be used as the gold-standard, user interest can be captured adaptively. Furthermore, because existing studies only examine review helpfulness at the document level, the helpfulness information is simply used as a filtering criteria (i.e. excluding unhelpful reviews when generating review summaries (Liu et al., 2007)) or ranking criteria (as seen on e-ecommerce websites). If we can have the ability to identify review helpfulness in finer granularities, other ways of using review helpfulness for summarization can be possible.

In this work, we investigate these opportunities based on a standard extractive multi-document summarization framework. In particular, we propose helpfulness-guided summarization methods which exploit review helpfulness at multiple levels of granularity: the proposed summarizers assess content informativeness not only by filtering but also by sentence-level review helpfulness predictions. Note that review helpfulness may be perceived in different ways from one domain to another, depending on the audience that the reviews mean to serve. Therefore, the performance of review summarization is supposed to be evaluated with target users. While there exist automatic (or semi-automatic) summarization evaluation methods, we rely mostly on user studies in evaluating our proposed summarizers.

Motivated by the opportunities and challenges in the research landscape of review helpfulness analysis and review summarization, the specific thesis work is further described below.

1.2 RESEARCH SUMMARY

The goal of this research is 1) to explore automatic review helpfulness prediction in general settings and 2) to enhance review summarization by leveraging review helpfulness at multiple levels of granularity. For the first one, we investigate two different ways to predict review helpfulness, one is based on augmenting an existing model with domain-specific information to capture domain-related semantics of review helpfulness; the other is by using a general review content representation that reflects review helpfulness in different domains. To achieve the second goal, we build new review summarizers by modifying an existing multi-document summarization framework to use review helpfulness for summarization content selection. We present two user studies to evaluate our helpfulness-guided summarization framework in different application scenarios with target users respectively.

To demonstrate the generality of our work, we experiment with our methods on three representative domains: one is most widely studied in the literature (product reviews (Jindal and Liu, 2008)); one is found most challenging for sentiment analysis (Turney, 2002) (movie reviews); another one is brand new (educational peer reviews (Xiong and Litman, 2011a)).

1.2.1 Review helpfulness prediction on educational peer reviews

Educational peer reviews have rarely been studied before. Prior work on review helpfulness analysis only considered customer reviews, in which various type of features are proposed for automatically predicting review helpfulness. Therefore, for predicting educational peer review helpfulness, our first solution is to tailor existing techniques found effective in traditional domains to the new peer-review domain, in which we hypothesize that:

1. Techniques used to predict review helpfulness in other domains can also be applied to educational peer reviews. (H1)
2. Incorporating peer-review domain knowledge as auxiliary features can improve prediction performance. (H2)

In Chapter 3, as a starting point, we refer to Kim's work on camera reviews (Kim et al., 2006) for experimental set up. We consider the helpfulness prediction task as a ranking problem that can be solved by supervised machine learning. To capture the educational semantics of peer-review helpfulness, we refer to empirical analysis in education and cognitive science to develop peer-review specialized features.

Our quantitative comparison shows that the utility of the features developed for customer reviews (generic features) in predicting review helpfulness varies between different review domains, while the proposed peer-review specialized features are predictive of peer-review helpfulness. Furthermore, we show that incorporating the peer-review specialized features with the generic features significantly improves the model's prediction performance.

1.2.2 New feature representation of review textual content for predicting review helpfulness across domains

While our work on educational peer reviews takes a specialization approach in designing the domain-specific auxiliary features, with respect to all kinds of reviews, we wonder if there is a general feature representation that predicts review helpfulness well independent of the review domain. Considering the generality of review information available across domains and the model's potential for being integrated into downstream applications such as review summarization, our second solution focuses on only using **review textual content** (language and semantics) for helpfulness prediction. We investigate review language style and expressiveness, as well as content diversity and topics, hoping to capture the semantics of review helpfulness in different domains.

For this purpose, we explore review content in two new (orthogonal) directions. One is about linguistic cues: we analyze review word use and language style based on existing dictionaries that categorize words with respect to their syntactic and semantic functions; we introduce statistical measurement of language expressiveness to describe review content diversity; we apply supervised statistical topic modeling to discover review latent topics associated with review helpfulness.

The other direction is based on splitting review content regarding where it comes from. To explain, consider the following example from an IMDB movie review on *Django Unchained (2012)*:

> *"Schultz tells Django to pick out whatever he likes. Django looks at the smiling white man in disbelief. You're gonna let me pick out my own clothes ? Django can't believe it.* The following shot delivered one of the biggest laughs from the audience I watched the film with. ..."

This review not only contains the reviewer's evaluation of the movie ("the following shot..."), but also contains a description of the movie plot (as italicized). While the evaluation is **the reviewer**'s opinion about the movie, the description is indirectly quoted from **the movie**. Distinguishing review-subject descriptions (or related content) and reviewers' evaluations may help us better predict review helpfulness. Although such non-homogeneity nature of review content has received little attention in prior work on review helpfulness, researchers in opinion mining have examined opinion sources (namely, opinion holders) in in news articles which often contain descriptions of opinions that do not belong to the article's author (Wiebe et al., 2005; Bethard et al., 2004; Choi et al., 2005). In contrast, our work does not focus on just opinions, instead, we examine the overall review content and propose an approximated differentiation method.

In Chapter 4, we analyze the predictive power of the new proposed features and the impact of review content types on review helpfulness across domains. In particular, we investigate the following hypotheses with the machine learning task of predicting review helpfulness:

3. Review helpfulness can be predicted using only review text, based on the same computational linguistic representation across domains. (H3)
4. The proposed high-level feature representation is more predictive of review helpfulness than low-level representation of review semantics (unigrams). (H4)
5. Distinguishing review-subject descriptions and other review content facilitates review helpfulness prediction. (H5)

While we find that the proposed content features are more predictive than simple unigrams for both the new domain of education and the challenging domain of movie reviews, unigrams still work best for product reviews. Despite that the most predictive content type and the corresponding features' utility vary with the domain, the proposed content splitting method constantly improves the prediction performance when using the combination of feature sets derived from both content types separately. More details will be discussed in Section 4.5.

1.2.3 Exploiting review helpfulness information in review summarization

In addition to automatically predicting review helpfulness, we also explore novel ways of utilizing review helpfulness in review summarization. Here we consider introducing review helpfulness at two levels of granularity. At the document level, we rely on review helpfulness gold standards, and use it to filter out unhelpful reviews (as how people did (Liu et al., 2007)). At the sentence level, we first learn a set of helpfulness-related latent topics from the review corpus as well as their utility scores in predicting the helpfulness at the document level. Then we infer the helpfulness of a sentence by aggregating the utility scores of the topics in the sentence. The sentence-level helpfulness predictions can then be used as features in summarization content selection. Our helpfulness-guided summarization approach is based on the following hypotheses:

6. User-provided review helpfulness assessment can be used to improve summarization performance. (H6)
7. Review helpfulness can be automatically predicted at the sentence level. (H7)
8. Using sentence-level review helpfulness information in addition to review-level helpfulness ratings yields better review summarizers. (H8)

We test these hypotheses by implementing our ideas based on an existing multi-document summarization framework. Specifically, we compare three review summarizers: 1) a nonhelpfulness baseline, 2) a summarizer that considers document-level helpfulness in addition to the features used in the baseline, and 3) a summarizer that considers review helpfulness at both the document and the sentence level, plus only relying on helpfulness features for summarization.

Since for camera reviews and movie reviews the target users can be anyone familiar with standard websites such as Amazon.com and IMDB.com, we conduct user studies with subjects recruited by email and social media. However, for educational peer reviews, the target users are students who receive the peer reviews. Therefore, we conduct a separate user study with students recruited in a Physics class at the University of Pittsburgh, using their received peer reviews.

Results show that our helpfulness-guided summarizers can outperform the baseline in both human and automated evaluation for both camera reviews and movie reviews. For educational peer reviews, the preference for helpfulness is significantly influenced by student writing performance and prior experience in teaching: low-performance students and non-expert students like the filtering approach but think the traditional summarizer more effective than the one using sentence-level helpfulness for content selection; high-performance and expert students think using both review-level and sentence-level helpfulness better than using review-level helpfulness alone in terms of content recall and accuracy.

1.3 CONTRIBUTIONS

Our research mainly contributes to review helpfulness prediction and review summarization.

First, our work successfully demonstrates that techniques used in predicting product review helpfulness can be effectively adapted to the domain of peer reviews and that peer-review domain knowledge can be further integrated by introducing new features that capture helpfulness information specific to peer reviews. This not only provides an empirical example of how to tailor existing techniques to a new domain, but also sheds light on developing automated peer-review assessment tools in computer based peer-review learning environments.

Second, we propose a set of general content features to predict review helpfulness in a general setting across domains only based on review (textual) content. The proposed content features do not depend on information of the review context (e.g., review star rating, reviewer profile), the review item (e.g., a particular product) or the review domain (e.g., product reviews vs. movie reviews). This general feature representation is also compact, reducing feature redundancy and over-fitting, leading to predictive models that significantly outperform traditional unigram-based lexical representations on challenging review domains such as movie reviews and peer reviews. Moreover, our helpfulness-related topic features support fine-grain analysis of review helpfulness within a review, which provides new opportunities for leveraging review helpfulness in applications such as summarization. In addition to the new content features, we demonstrate the importance of content categorization regarding its reference to review subject before feature engineering for review helpfulness prediction.

Finally, we propose a novel unsupervised extractive approach for summarizing online reviews by exploiting user-provided review helpfulness for sentence-scoring in summarization content selection. We demonstrate that document-level helpfulness can not only be directly used for review filtering, but also be used to infer sentence-level review helpfulness features for sentence scoring. This approach leverages the existing metadata of online reviews, requiring no annotation and generalizing to multiple review domains. In a broader view, our work provides evidence for taking into account review helpfulness in review summarization, which suggests similar consideration of helpfulness information for other review-related tasks. Meanwhile, our work suggests a promising solution of adapting standard multi-document summarization techniques for building educational applications, which also contributes to communities outside of NLP, such as educational data mining (EDM) and intelligent tutoring systems (ITS), etc.

1.4 THESIS OUTLINE

The rest of the thesis is organized in the following way:

In Chapter 2, we introduce the three review corpora that we use in the presented studies (including examples of helpful versus unhelpful reviews), as well as explain the gold-standard that we used for supervised machine learning experiments.

In Chapter 3, we tackle the problem of automatically assessing the helpfulness of educational peer reviews. We examine prior techniques that have been used to successfully rank helpfulness for product reviews, and evaluate whether these techniques also apply to our new context of peer reviews. Furthermore, we investigate the utility of incorporating additional specialized features tailored to peer review.

In Chapter 4, motivated by downstream applications such as review summarization, we consider modeling review helpfulness in a more general perspective using only review text and experimenting within three distinct review domains. We propose features to represent review language usage, content diversity and helpfulness-related topics along with a content categorization method used as a preprocessing step, and compare different models in machine learning experiments.

In Chapter 5, we explore how to utilize review helpfulness for review summarization by modifying a standard multi-document summarization framework. We present two user studies to illustrate the merits of our proposed helpfulness-guided summarizers against a standard multi-document summarizer: one is summarizing customer reviews for a camera or a movie, the other is summarizing peer reviews for students in an educational setting.

Finally in Chapter 6, we summarize the major discoveries that we obtained in our work, and highlight our contributions to the related communities.

2.0 DATA SETS

To demonstrate the generality of the proposed ideas, we consider three distinct review domains throughout the thesis: product reviews, movie reviews and peer reviews.

Electronic product reviews were the first kind of online review studied in the area of review helpfulness and thus are frequently included in later work. Movie reviews have also been studied and seem to have more diverse content. While the emphasis of a product review is usually the product (e.g., the camera), the emphasis of a movie review can diverge from the movie plot to the reviewer's personal thoughts on moral, social, and ethical issues. Peer reviews, a much newer kind of user review, serves a different function due to their educational context. Both movie reviews and peer reviews are potentially more complicated than the product reviews, as the review content consists of both the reviewer's evaluations of the subject (e.g., a movie or paper) and the reviewer's references of the subject, where the subject itself is full of content (e.g., movie plot, papers). In contrast, such references in product reviews are usually the mentions of product components or properties, which have limited variations. This characteristic makes review summarization more challenging in these two domains.

2.1 PRODUCT REVIEWS

For product reviews, we use Amazon review data provided by Jindal and Liu (2008), as this is a widely used data set in review opinion mining and sentiment analysis. In particular, we choose one representative product type, digital camera, which is one category of the products that have been most widely studied.

In this data set, various metadata is also available for each review, such as the product name/ID, product rating, the total number of helpfulness votes, the number of "helpful" votes, etc. The helpfulness vote is a binary vote of a review being "helpful" vs. "unhelpful". Online readers can vote for the helpfulness of the product reviews through Amazon.com. To ensure the quality of user-provided helpfulness assessment, we filter out reviews that were voted on by less than 3 people. In total, there are 4050 camera reviews. Descriptive statistics are summarized in Table 2.1.

Measurement	**Camera**	**Movie**	**History2008**	**Physics2014**
Vocabulary size	13160	9492	2699	4996
# of reviews	4050	280	267	6203
# of words / review	170	435	101	34
average helpfulness	.80	.74	.43	.84

Table 2.1: Descriptive statistics of review corpora used in this thesis.

Camera review examples: *Canon PowerShot SD600 6MP Digital Elph Camera with 3x Optical Zoom*

- A more helpful camera review: (38 out of 38 people voted "yes" for the helpfulness of the review)

> *Funny how this camera seems to be tested in Boston. I bought this camera specifically for a trip to Boston.*
> *I thought my brother's SD450 was the best digital camera I've ever used, until I got the SD600. I took TONS of pictures in different lighting and all turned out great. I shot everything from plates of food inches away to soccer games with players clear across the field.*
> *Certain settings worked better than others for different lighting situations. It really just depends on what color tones you prefer. I prefer warmer tones and the camera worked really well. For the basic point-and-shoot shots with the setting on automatic and flash, I thought the colors were very true to life...*

- A less helpful camera review: (5 out of 46 people voted "yes" for the helpfulness of the review)

> *After waiting months for an order to be filled and then getting screwed by customer service/returns, I decided never to buy from Amazon again. Go to an electronics or camera store, you'll get better service and the fresh air will do you good.*

2.2 MOVIE REVIEWS

Our movie review data set is crawled by ourselves from IMDB.com because the helpfulness metadata is not available in existing movie review corpora that we are aware of (till December 2013). To create this data set, we pick 7 famous movies[1] and collect the top 40 user reviews based on their helpfulness for each one of them, including the corresponding helpfulness votes, movie keywords, plot summaries, synopses, etc. By default, user reviews are displayed in the order of their helpfulness on IMDB.com. However, the helpfulness scoring function used by IMDB.com is different from ours in that it also considers the total number of helpfulness votes (e.g., a review that 6 out of 9 users voted as helpful is ranked

[1] The movies are: "The Godfather", "The Dark Knight", "De-Lovely", "The Lord of The Ring, The Return of the King", "Pulp Fiction", "Forrest Gump" and "Shawshank Redemption".

higher than one that 2 out of 2 users voted as helpful.). We stick with the default ranking, as it allows us to prepare a movie review corpus with helpfulness rating statistics similar to those of the product reviews and in which each review has at least certain number of votes.

To make our movie review corpus comparable to our peer review corpus (History2008, which is introduced later), we only prepare 280 movie reviews. We use the movie keywords, plot summaries and synopses to identify the review content that is about the movie itself (Section 4.5.2). Descriptive statistics about this data set are included in Table 2.1.

Movie review examples: *Forrest Gump*

- A more helpful movie review: (121 out of 142 people voted "yes" for the helpfulness of the review)

> *"I've made about 20 films and 5 of them are pretty good" -Tom Hanks. "Forrest Gump" is one of the best movies of all time, guaranteed. I really just love this movie and it has such a special place in my heart. The performances are just so unforgettable and never get out of your head. The characters, I mean the actors turned into them and that's what got to me. The lines are so memorable, touching, and sometimes hilarious.*
>
> *We have Forrest Gump (Tom Hanks), not the sharpest tool in the box, his I.Q. is right below the average scores. But his mama (Sally Field) believes that her boy has the same opportunities as anyone else and lets Forrest know that there's nothing that could hold him back. As a boy he is put into braces for his legs since he has a crooked back and really doesn't have too many friends. When he gets on a school bus for his first day of school, NO ONE will let him sit next to them. This scene is so heart breaking until you hear a little angel's voice "You can sit here if you want"...*

- A less helpful camera review: (15 out of 28 people voted "yes" for the helpfulness of the review)

> *All that money, all those clever effects, all those stars... and for what? A mind-numbing stream of syrup with no discernible purpose except to fool very dull people into thinking they have seen an epic movie. Life is like a box of chocolates? No, but some movies are as sickly sweet.*

2.3 EDUCATIONAL PEER REVIEWS

There are two peer-review data sets involved in this thesis; both are collected from an online peer-review reciprocal system (SWoRD (Cho, 2008)) developed at the University of Pittsburgh.

One data set (History2008) was collected in a college level history class, and has been annotated in a previous study (Nelson and Schunn, 2009) (Section 2.3.2.1). We use this data set for our studies of review helpfulness analysis (Chapter 3 and Chapter 4).

The other data set (Physics2014) was from a recent Physics Lab class at the University of Pittsburgh. This data set is only collected for evaluating our proposed summarization framework on educational peer reviews (conducted in spring 2014, Section 5.5).

2.3.1 Peer review using SWoRD

Here we briefly explain how educational peer review is done through SWoRD, using History2008 as an example:

Phase1. Assignment creation.

The teacher first created the writing assignment in SWoRD and provided a peer-review rubric that required students to assess a paper's quality on three separate dimensions (Logic, Flow and Insight), by giving a numeric rating on a scale of 1-7 in addition to textual comments. For instance, the teacher created the following guidance for commenting on the "Logic" dimension: "Provide specific comments about the logic of the author's argument. If points were just made without support, describe which ones they were. If the support provided doesn't make logical sense, explain what that is. If some obvious counter-argument was not considered, ..." Teacher guidance for numerically rating the logical arguments of the paper was also given. For this assignment, a rating of 7 ("Excellent") was described as "All arguments strongly supported and no logical flaws in the arguments". A rating of 1 ("Disastrous") was described as "No support presented for any arguments, or obvious flaws in all arguments".

Phase 2. Paper writing & peer review.

In the next phase, 24 students submit their papers online through SWoRD and then review (roughly) 6 peers' papers. The peer review is done in a "double blind" manner and each paper is reviewed by about 6 peers. As students are required to submit reviews on each dimension separately, SWoRD automatically associates the reviewing dimension with every numerical rating and textual comments.

Phase 3. Back-evaluation of review helpfulness.

Finally, the reviewers are rated backwards for their review helpfulness (at the review level on a scale of 1-7), by students who receive their reviews. We refer to these ratings as student-helpfulness ratings.[2]

Note that, peer reviews from different disciplines should be considered as in different sub-domains. Here we treat peer reviews collected from each assignment as a separate data set. In contrast with Amazon reviews, the peer review data sets are generally much smaller due to the limited number of students involved in a particular peer-review assignment.

2.3.2 Peer-review data: History2008

The first peer-review data set contains a paper corpus (24 student papers) and a review corpus of 267 peer reviews, generated from the peer-review activities described above (Section 2.3.1). In the prior work (Nelson and Schunn, 2009), a writing expert and a content expert also rated the review helpfulness for research purposes, which is further explained in Section 2.3.2.1. Descriptive statistics of the data set are included in Table 1. As it shows, peer reviews are different from product reviews in terms of the average number of reviews per subject and the average number of sentences per review.

Peer review examples *(History2008)*

[2] Also, the peer assessment algorithms in SWoRD will automatically evaluate reviewers' reviewing performance in terms of reviewing accuracy in a scale from 0 to 1.

- A helpful peer review: (both experts rate it as 5 points)

> *The support and explanation of the ideas could use some work. broading the explanations to include all groups could be useful. My concerns come from some of the claims that are put forth. Page 2 says that the 13th amendment ended the war. is this true? was there no more fighting or problems once this amendment was added? ...*
> *The arguments were sorted up into paragraphs, keeping the area of interest clear, but be careful about bringing up new things at the end and then simply leaving them there without elaboration (ie black sterilization at the end of the paragraph).*

- An unhelpful peer review: (both experts rate it as 1 point)

> *Your paper and its main points are easy to find and to follow.*

2.3.2.1 Manual annotations In a prior study of feedback utilities regarding revision likelihood, two experts (one domain expert and one content expert) examined the helpfulness of some of the peer reviews. They rated the helpfulness on a scale of 1-5 (Pearson correlation $r = .425$, $p < .01$). We consider the average rating given by the two experts for each review as the expert-helpfulness rating. In addition, all reviews were manually segmented into self-contained idea units (named as feedback), each of which was manually labeled for various properties and cognitive constructs that significantly correlate with the feedback implementation likelihood (Nelson and Schunn, 2009).

- feedbackType ($Kappa = .92$)

  The type of the peer review feedback. This property was coded with three values: *problem*, *praise* or *summary*. Only idea units that were coded as *problem* were coded for *problem localization* and *solution*.

- Problem localization ($Kappa = .69$)

  Whether the review feedback pinpoints to a specific place where the problem occurs in the paper. This construct is labeled as “True” or “False”, only for feedback that are coded as *problem*. pLocalization indicates whether the *problem* feedback contains *problem localization* for any specified problems.

- Solution ($Kappa = .79$)

  Whether the review feedback suggests any solution. This construct is labeled as "True" or "False", only for feedback that are coded as *problem*. Solution indicates whether the problem feedback provides actionable revision suggestions.

These kappa values (Nelson and Schunn, 2009) were calculated from a subset of the corpus for evaluating the reliability of human annotations, and these annotators are not the same experts who rated the peer-review helpfulness.

Annotation example:

In the following example, there are two idea units, each one is a separate paragraph. The annotations given to the first idea unit are: $feedbackType = problem$, $pLocalization = True$, $solution = True$. The annotations for the second one are: $feedbackType = problem$, $pLocalization = False$, $solution = True$.

> *I thought there were some good opportunities to provide further data to strengthen your argument. For example the statement "These methods of intimidation, and the lack of military force offered by the government to stop the KKK, led to the rescinding of African American democracy." Maybe here include data about how the percent of black people who voted based on the number of black people who were allowed to vote was extremely low. Specific numbers here would help your case, and there are other spots that the same can be said.*
>
> *Good use of citing your sources, but do yourself a favor and Bold all the citations you used and make it say 20 font. Print out a copy of the essay like that and then look at it. You'll notice that almost every line is taken from a source. I'm not saying that it wasn't nicely done, but on mere impressions this doesn't look good. It makes it look like there was no independent thought/argument and its all someone else's ideas.*

2.3.3 Peer-review data: Physics2014

This second peer review corpus was collected from a Physics class at the University of Pittsburgh, which used SWoRD for peer-review lab reports. It contains 6203 reviews and student provided helpfulness ratings. Compared with History2008, the reviews in this data set are much shorter, and the average helpfulness rating given by the students is much higher than the average helpfulness rating given by the experts in History2008, as indicated in Table 2.1.

As Physics2014 is especially prepared for our summarization evaluation, it does not have manual segmentation and annotations as History2008 has. So we did not include this data set in our review helpfulness prediction experiments. Peer-review examples from this data set are provided below. Note that the helpfulness ratings attached to the examples are given by students during peer-review Phase 3. In the particular examples provided below, both reviews are commenting on the same paper. The peer who wrote the first review rated the paper as 4 points, the one who wrote the second review rated it as 2 points.

Peer review examples *(from Physics2014)*

- A helpful peer review: (student rates it as 5 points)

> *All you really did was name the experiment and present Ohm's law. State what you are doing in this experiment and provide a quick summary and conclusions generated from it.*

- An unhelpful peer review: (student rates as 1 point)

> *Overall, the paper had correct grammar and spelling. However, try to get your writing to flow a little more. The introduction illustrates this well. Try to not just list concepts, steps, or results. Connect ideas and re-word sentences and paragraph structure to create a report that can be more pleasant to read.*

2.4 GOLD STANDARD OF REVIEW HELPFULNESS RATINGS

The definition of review helpfulness varies with the review domain, depending on why people read the reviews: to make a purchase decision (camera reviews), to pick a movie to watch (movie reviews), to know how to revise their papers (peer reviews), etc. The difference in the service that different kinds of reviews provide motivates us to explore whether/what aspects of review content reflect user-provided helpfulness ratings across domains. It is also the reason why we propose a general review helpfulness model just based on review text.

We follow prior studies of product reviews, defining camera and movie review helpfulness as the percentage of "helpful" votes that a review receives across multiple readers[3]. While prior studies on these online helpfulness votes point out that they are likely to be biased in various ways (e.g., helpful reviews are likely to get more votes), we leave further analysis on the robustness of our gold standard as future work.

For educational peer reviews, prior work shows that how students rate the helpfulness of the reviews that they receive from their peers depends on how these peers rated their papers. This makes student-helpfulness ratings less validated compared to expert-helpfulness ratings, as suggested in (Xiong and Litman, 2011b). Therefore, we consider expert-helpfulness ratings (Nelson and Schunn, 2009) (History2008) as the helpfulness gold standard in educational peer reviews for building our automated helpfulness assessment models in Chapter 3 and Chapter 4. But when using helpfulness for review summarization, we use student-helpfulness ratings (Physics2014), as they are "user provided helpfulness assessment". (Students are the target users of our review summarizers in the education domain.)

For consistency, we scale all helpfulness assessment in the range between 0 and 1. In our experiments, we will use these normalized ratings as the gold standard of review helpfulness.

The descriptive statistics are summarized in Table 2.1.

[3] Multiple users provide helpfulness votes ("helpful" vs. "unhelpful") for reviews on Amazon.com and IMDB.com

3.0 AUTOMATICALLY PREDICTING PEER-REVIEW HELPFULNESS

This chapter explores the feasibility of applying standard review analysis techniques to a new review domain – educational peer reviews.[1] In particular, we consider review helpfulness prediction as a supervised machine learning problem. We examine the effectiveness of the features that are proposed for predicting product review helpfulness in the context of peer reviews. In addition, we investigate the utility of incorporating additional auxiliary features that are specific to peer review.

Our results show that structural features, review unigrams and metadata combined are useful in modeling the helpfulness of both peer reviews and product reviews, while peer-review specific auxiliary features can further improve helpfulness prediction.

3.1 RELATED WORK

Prior studies of peer review in the Natural Language Processing field have not focused on helpfulness prediction, but instead have been concerned with issues such as highlighting key sentences in papers (Sandor and Vorndran, 2009), detecting important feedback features in reviews (Cho, 2008; Xiong and Litman, 2010), and adapting peer-review assignment (Garcia, 2010). However, given some similarity between peer reviews and other review types, we hypothesize that techniques used to predict review helpfulness in other domains can

[1] This piece of work was published in ACL2011 as a short paper (Xiong and Litman, 2011a).

also be applied to peer reviews.

Early work (Kim et al., 2006) on Amazon's product reviews showed that review helpfulness can be automatically predicted using features derived from review text and review metadata. Kim et al. examined various types of structural, lexical, syntactic, semantic and meta features and compare their utility for predicting review helpfulness using SVM regression. While sentiment words and domain lexicons (semantic features) were computed, experimental results showed that unigrams were the most predictive features and would subsume other syntactic and semantic features. Ghose and Ipeirotis (2011) used a similar approach to examine the socio-economic impact and the perceived helpfulness of product reviews. They suggested the usefulness of subjectivity analysis, reviewer information, and review readability features, though they also found that their predictive power is interchangeable. In addition, other studies showed that the perceived helpfulness of a review depends not only on its review content, but also on social effects such as product qualities, and individual bias in the presence of mixed opinion distribution (Danescu-Niculescu-Mizil et al., 2009). Considering that Kim et al.'s work has a broader coverage of feature types, we use their feature set as our basis for predicting peer review helpfulness. As different features have been proposed for different types of reviews, we only focus on product reviews in this survey and leave the other domains to the next chapter where we will survey how review content has been modeled using different types of features in all kinds of reviews.

With respect to the learning techniques, the helpfulness prediction task has been modeled in different ways. While many studies defined helpfulness ratings as a numeric helpfulness variable (e.g., the percentage of "yes" votes), some researchers converted the helpfulness measurement into a categorical variable (usually binary). In the former case, helpfulness prediction was usually considered as a ranking problem that can be solved by regression; in the latter case, helpfulness prediction was treated as a classification task. Unsupervised learning has also been explored. Tsur and Rappoport (2009) proposed RevRank to select the most helpful book reviews based on the similarity between the review content and a set of dominant lexicons that were identified from the whole review corpus. In comparison, the supervised methods require review helpfulness gold-standards, though different gold-standards could be used which might impact the features used to build the helpfulness model. For example, Liu et al. (2007) used their own annotation of review helpfulness for Amazon product reviews based on a set of specifications based on the coverage and the level of detail of the review content; a lot of their features are about recognizing mentions of product names and properties. The unsupervised methods consider helpful reviews as those that contain dominant content, however, some under-represented ideas could be helpful too. In our studies, we follow the majority of the prior work, using existing supervised ranking algorithms to predict review helpfulness as a numeric variable.

In terms of domain differences, several properties distinguish our corpus of peer reviews from other types of reviews: 1) The helpfulness of our peer reviews is directly rated using a discrete scale from one to five instead of being defined as a function of binary votes (e.g., the percentage of "helpful" votes (Kim et al., 2006)); 2) Peer reviews frequently refer to the related students' papers, thus review analysis needs to take into account paper topics; 3) Within the context of education, peer-review helpfulness often has a writing specific semantics, e.g. improving revision likelihood; 4) In general, peer-review corpora collected from classrooms are of a much smaller size compared to online product reviews. To tailor existing techniques to peer reviews, we will thus propose new specialized features to address these issues.

In this work, we mainly refer to empirical studies of educational peer reviews in cognitive science for developing peer-review specialized features. In the analysis of History2008, Nelson and Schunn (2009) found that several cognitive constructs are predictors of whether the review comments (problems) were addressed in student future revisions. Among them, 1) the presence of localization information regarding where the problem occurred (*problem localization*), and 2) concrete solutions provided to address the problem (*solution*). Here we propose computational linguistic features to capture the patterns of the cognitive constructs in peer reviews, and compare their utility in prediction review helpfulness with features derived directly from human labels of these constructs.

3.2 FEATURES

Our features are motivated by the prior work introduced above, in particular, NLP work on predicting product-review helpfulness (Kim et al., 2006), as well as work on automatically learning cognitive-science constructs (Nelson and Schunn, 2009) using natural language processing (Cho, 2008; Xiong and Litman, 2010).

All the computational linguistic features are automatically extracted based on the output of syntactic analysis of reviews and papers[2].

3.2.1 Generic features

We first mine **generic** linguistic features from reviews and papers, aiming to replicate the feature sets used by Kim et al. (2006). These generic features are summarized in in Table 3.1.

While structural, lexical and syntactic features are created in the same way as suggested in (Kim et al., 2006), we adapt the semantic and meta-data features to peer reviews 1) by converting the mentions of product properties to mentions of the history topics, and 2) by using paper ratings assigned by peers instead of product scores.

- **Structural features** consider the general structure of reviews, which includes review length in terms of tokens, number of sentences, the average sentence length, percentage of sentences that end with question marks, and number of exclamatory sentences.
- **Lexical features** include review unigrams and bigrams, where each term is weighted by their tf-idf score.

[2] We used MSTParser (McDonald et al., 2005) for syntactic analysis.

- **Syntactic features** mainly focus on nouns and verbs, and include percentage of tokens that are nouns, verbs, verbs conjugated in the first person, adjectives/adverbs, and open classes, respectively.

- **Semantic features** capture two important peer-review properties: their relevance to the main topics in students' papers, and their opinion sentiment polarities. Kim et al. (2006) extracted product property keywords from *external resources* based on their hypothesis that helpful product reviews refer frequently to certain product properties. Similarly, we hypothesize that helpful peer reviews are closely related to domain topics that are shared by all student papers in an assignment. Our domain topic set contains 288 words extracted from the collection of student papers, which is the *external resource* of the peer reviews, using topic-lexicon extraction software[3]; our feature (*TOP*) counts how many words of a given review belong to the extracted set. For sentiment polarities, we extract positive and negative sentiment words from the General Inquirer Dictionaries [4], and count their appearance in reviews in terms of their sentiment polarity (*posWord, negWord*). While we acknowledge that there are other sentiment lexicon dictionaries available to use (e.g., MPQA (Wiebe et al., 2005)), we picked the same one that Kim et al. used in their work (Kim et al., 2006).

- **Metadata features** are derived from student paper ratings to reflect interactions between students in a peer-review assignment. As suggested in (Kim et al., 2006), some social dimensions (e.g., customer opinion on related product quality) are of great influence in review helpfulness. We similarly take the social aspects of peer review into account by introducing related paper ratings and the absolute difference between the rating and the average score given by all reviewers who reviewed the same paper.

[3] The software extracts topic words based on topic signatures (Lin and Hovy, 2000), and was kindly provided by Annie Louis (http://homepages.inf.ed.ac.uk/alouis/topicS.html).

[4] http://www.wjh.harvard.edu/ inquirer/homecat.htm

Class	Label	Features
Structural	STR	review length in terms of tokens, number of sentences, the average sentence length, percentage of sentences that end with question marks, number of exclamatory sentences.
Lexical	UGR, BGR	*tf-idf* statistics of review unigrams and bigrams.
Syntactic	SYN	percentage of tokens that are nouns, verbs, verbs conjugated in the first person, adjectives / adverbs and open classes, respectively.
Semantic	TOP, posWord, negWord	counts of topic words, counts of positive and negative sentiment words.
Metadata	MET	the overall ratings of papers assigned by reviewers, and the absolute difference between the rating and the average score given by all reviewers.

Table 3.1: Generic features motivated by related work of product reviews.

3.2.2 Peer-review specialized features

In addition, the following new peer-review **specialized** features are developed to model review helpfulness in the educational context, motivated by Nelson and Schunn (2009). Among these features, *lexCat* and *LOC* are computational linguistic features aiming to capture the important cognitive constructs in peer reviews (introduced in Section 3.1), while *cogSci* are directly derived from the manual annotations of those constructs (introduced in Chapter 2).

- **Lexical category features (lexCat).** We first take the advantage of our domain expertise and crafted a table of keywords (Table 3.2) to captures the lexical signals in

Tag	Meaning	Word list
SUG	suggestion	should, must, might, could, need, needs, maybe, try, revision, want
LOC	location	page, paragraph, sentence
ERR	problem	error, mistakes, typo, problem, difficulties, conclusion
IDE	idea verb	consider, mention
LNK	transition	however, but
NEG	negative words	fail, hard, difficult, bad, short, little, bit, poor, few, unclear, only, more
POS	positive words	great, good, well, clearly, easily, effective, effectively, helpful, very
SUM	summarization	main, overall, also, how, job
NOT	negation	not, doesn't, don't
SOL	solution	revision specify correction

Table 3.2: Ten lexical categories.

important cognitive constructs, such as *feedbackType* and *Problem localization*, etc. As no existing dictionary is specialized for peer-review analysis, we constructed our own lexical categories with manual editing to avoid errors that might be introduced by any automated methods.

To construct the keyword table, two domain experts first manually selected a set of keywords from the instructions[5] on annotating *feedbackType*, *Problem localization* and *Solution* (independent of the peer-review domain), and categorized them according to the semantics of the keywords. Then we trained a decision tree for *feedbackType* clas-

[5] This is the annotation guide used in (Nelson and Schunn, 2009).

sification based on review unigrams (bag-of-words), using other annotated peer-review corpora[6]. We manually examined the unigrams that were selected to form the decision tree and hand-picked ones independent of the peer-review domain to supplement the lexical categories. These categories are shown to be effective in automatic *feedbackType* identification (Xiong et al., 2010).

- **Localization features (LOC)** are developed to capture the pattern of *problem localization* in particular, as *problem localization* is found to be most influential in peer-review feedback implementation (Nelson and Schunn, 2009). To be specific, the localization features are constructed as the following:

  1. Simple regular expressions (RE) are first employed to recognize common location phrases such as "on page 5" and "the section about". We check each "problem" sentence to see if any RE is matched, and then calculate the percentage of the "problem" sentences that are matched to at least one RE as one LOC feature.
  2. The syntactic structure of review sentences is considered as well. We check whether there is any domain topic word[7] between the subject and the object in any sentence, and also count demonstrative determiners in each review sentence. Then we calculate the percentage of sentences that has at least one domain topic word between its subject and its object, and the average number of demonstrative determiners per sentence as part of the LOC features.
  3. The features above are based on our intuition about localized expressions, while the following ones are derived from an overlapping-window algorithm that was shown to be effective in a similar task – identifying quotation from reference works in primary materials for digital libraries (Ernst-Gerlach and Crane, 2008). To match a possible citation in a reference work, it searches for the most likely referred window of words through all possible primary materials. We applied this algorithm

[6] The other annotated peer review corpora are in Physics and Cognitive Science. They were used as development data sets in our pilot study, which is separate from the peer review data set presented in this work

[7] We construct the domain topic set when creating the semantic feature TOP in Section 3.2.1.

for our purpose, and consider the average length of the window plus the average number of words of all windows.

To illustrate how these features are computed, consider the following critique:

> *The section of the essay on African Americans needs more careful attention to the timing and reasons for the federal governments decision to stop protecting African American civil and political rights.*

The review has only one sentence, in which one regular expression is matched with "the section of" thus $regTag\% = 1$; no demonstrative determiner, thus $dDeterminer = 0$; "African" and "Americans" are domain words appearing between the subject "section" and the object "attention", so *soDomain* is true for this sentence and thus $soDomain\% = 1$ for the given review.

The LOC features have also been used to build a classifier for identifying peer-review *Problem localization*, and the corresponding work is published in (Xiong and Litman, 2010).

- **Cognitive-science features (cogSci).** To examine how the computational linguistic features above capture the cognitive constructs in the context of peer-review helpfulness prediction, we compare them with a third type of specialized features directly from human labels of these constructs. Therefore, cogSci can be considered as an upper bound of the performance of our automated features for capturing the important cognitive constructs.

  In our data set, the cognitive-science constructs are manually coded at the level of idea unit (self-contained text span) (Nelson and Schunn, 2009), however, the peer-review helpfulness is rated for the whole review, which can include multiple idea units.[8] Therefore in our study, we calculate the distribution of *feedbackType* values (*praise*, *problem*, and *summary*), the percentage of problems that are *localized* ($problemlocalization = True$), and the percentage of problems that have a *solution*

[8] Details of different granularity levels of annotation can be found in (Nelson and Schunn, 2009).

($solution = True$) to model peer-review helpfulness.

> *(Unit 1) The support and explanation of the ideas could use some work. Broading the explanations to include all groups could be useful. My concerns come from some of the claims that are put forth. Page 2 says that the 13th amendment ended the war. is this true? was there no more fighting or problems once this amendment was added? ...*
>
> *(Unit 2) The arguments were sorted up into paragraphs, keeping the area of interest clear, but be careful about bringing up new things at the end and then simply leaving them there without elaboration (ie black sterilization at the end of the paragraph).*

Consider the review example above, which was manually separated into two idea units (each presented in a separate paragraph). As both ideas are coded as *problem*, $problemlocalization = True$ and $solution = True$, the cognitive-science features of this review are *praise%*=0, *problem%*=1, *summary%*=0, *localization%*=1, and *solution%*=1.

3.3 EXPERIMENTAL SETUP

The following experiment is designed to verify our hypotheses about the specialization approach for review helpfulness prediction:

H1 Techniques used to predict review helpfulness in other domains can also be applied to educational peer reviews.

H2 Incorporating peer-review domain knowledge as auxiliary features can improve prediction performance.

In this experiment, we use the previously annotated peer-review corpus History2008. Recall that the corpus consists of 16 papers (about six pages each) and 267 reviews (varying from twenty words to about two hundred words). As two experts rated the helpfulness of each peer review on a scale from one to five with respect to content and writing independently (Pearson correlation $r = 0.425$, $p < 0.01$) (Nelson and Schunn, 2009), we consider the average ratings given by the two experts (which roughly follow a normal distribution) as the gold standard of review helpfulness.

As we choose Kim et al. (2006)'s work as the basis to develop our peer review feature set, we follow their work and train our helpfulness model using SVM regression with a radial basis function kernel provided by SVM^{light} (Joachims, 1999). To test Hypothesis **H1**, we first evaluate each feature type in isolation to investigate its predictive power of peer-review helpfulness; to test Hypothesis **H2**, we then examine them together in various combinations to find the most useful feature set for modeling peer-review helpfulness.

Performance is evaluated in 10-fold cross validation of our 267 peer reviews by predicting the absolute helpfulness scores (with Pearson correlation coefficient r) as well as by predicting helpfulness ranking (with Spearman rank correlation coefficient r_s). Although predicted helpfulness ranking could be directly used to compare the helpfulness of a given set of reviews, predicting helpfulness rating is desirable in practice to compare helpfulness between existing reviews and new written ones without reranking all previously ranked reviews. Results are presented regarding the generic features and the specialized features respectively, with 95% confidence bounds.

3.4 RESULTS

3.4.1 Performance of generic features

Evaluation of the generic features is presented in Table 3.3, showing that all classes except syntactic (SYN) and meta-data (MET) features are significantly correlated with both helpfulness rating (r) and helpfulness ranking (r_s). Structural features (bolded) achieve the highest Pearson (0.60) and Spearman correlation coefficients (0.59) (although within the significant correlations, the difference among coefficients are insignificant). Note that in isolation, MET (paper ratings) are not significantly correlated with peer-review helpfulness, which is different from prior findings of product reviews (Kim et al., 2006) where product scores are significantly correlated with product-review helpfulness. However, when combined with other features, MET does appear to add value (last row). When comparing the performance between predicting helpfulness ratings versus ranking, we observe $r \approx r_s$ consistently for our peer reviews, while Kim et al. (2006) reported $r < r_s$ for product reviews.[9] Finally, we observed a similar feature redundancy effect as Kim et al. (2006) did, in that simply combining all features does not improve the model's performance. Interestingly, our best feature combination (last row) is the same as theirs. In sum our results verify our hypothesis that the effectiveness of generic features can be transferred to our peer-review domain for predicting review helpfulness.

3.4.2 Analysis of the peer-review specialized features

Evaluation of the specialized features is shown in Table 3.4, where all features examined are significantly correlated with both helpfulness rating and ranking. When evaluated in isolation, the computational linguistic features (lexCat and LOC) outperform the human-label based features (though the difference is not significant). Although specialized features have weaker correlation coefficients ([0.43, 0.51]) than the best generic features (.06), these differences are not significant, and the specialized features have the potential advantage of being theory-based. The use of features related to meaningful dimensions of writing has contributed to validity and greater acceptability in the related area of automated essay scoring (Attali and Burstein, 2006).

When combined with some generic features, the specialized features improve the model's

[9] The best performing single feature type reported (Kim et al., 2006) was review unigrams: $r = 0.398$ and $r_s = 0.593$.

Features	**Pearson r**	**Spearman r_s**
STR	**0.60(0.10)***	**0.59(0.10)***
UGR	0.53(0.09)*	0.54(0.09)*
BGR	0.58(0.07)*	0.57(0.10)*
SYN	0.36(0.12)	0.35(0.11)
TOP	0.55(0.10)*	0.54(0.10)*
posWord	0.57(0.13)*	0.53(0.12)*
negWord	0.49(0.11)*	0.46(0.10)*
MET	0.22(0.15)	0.23(0.12)
All-combined	0.56(0.07)*	0.58(0.09)*
STR+UGR+MET+TOP	0.61(0.10)*	0.61(0.10)*
STR+UGR+MET	**0.62(0.10)***	**0.61(0.10)***

Table 3.3: Performance evaluation of the generic features for predicting peer-review helpfulness. Significant results are marked by * ($p \leq 0.05$).

performance in terms of both r and r_s compared to the best performance in Table 3.3 (the baseline). Though the improvement is not significant yet, we think it still interesting to investigate the potential trend to understand how specialized features capture additional information of peer-review helpfulness. Therefore, the following analysis is also presented (based on the absolute mean values), where we start from the baseline feature set, and gradually expand it by adding our new specialized features: 1) We first replace the raw lexical unigram features (UGR) with lexical category features (lexCat), which slightly improves the performance before rounding to the significant digits shown in row 5. Note that the categories not only substantially abstract lexical information from the reviews,

Features	**Pearson r**	**Spearman r_s**
Lexical categories (lexCat)	0.51(0.11)	0.50(0.10)
Localization (LOC)	0.45(0.13)	0.47(0.11)
Cognitive science (cogSci)	0.43(0.09)	0.46(0.07)
STR+MET+UGR (Baseline)	0.62(0.10)	0.61(0.10)
STR+MET+lexCat	0.62(0.10)	0.61(0.09)
STR+MET+lexCat+TOP	0.65(0.10)	0.66(0.08)
STR+MET+lexCat+TOP+LOC	0.65(0.09)	0.66(0.08)
STR+MET+lexCat+TOP+LOC+cogSci	**0.67(0.09)**	**0.67(0.08)**

Table 3.4: Evaluation of the model's performance (all significant) after introducing the specialized features.

but also carry simple syntactic and semantic information. 2) We then add one semantic class – topic words (row 6), which enhances the performance further. Semantic features do not help when working with generic lexical features as shown in Table 3.3 (second to last row), but they can be successfully combined with the lexical **category** features and further improve the performance as indicated here. 3) When LOC is further added (row 7), the performance is maintained, with a Pearson correlation of 0.65 and a Spearman correlation of 0.66. 4) But we also notice that the automated features have not yet fully represented the cognitive science constructs, as adding human-label based features can achieve slightly better performance (Table 3.4, last row). However, in real operational settings when the cogSci features are not available, the computational linguistic features can be used to achieve comparable prediction performance (as shown in row 7).

In sum, we confirm our hypotheses that existing review helpfulness prediction methods developed for product reviews can be tailored to educational peer reviews, and using our proposed peer-review specialized features in combination with the generic features increases our model's predictive power.

3.5 DISCUSSION

Despite the difference between peer reviews and other types of reviews as discussed in Section 3.1, our work demonstrates that many generic linguistic features are also effective in predicting peer-review helpfulness. The model's performance can be alternatively achieved and further improved by adding auxiliary features tailored to peer reviews. These specialized features not only introduce domain expertise, but also capture linguistic information at an abstracted level, which can help avoid the risk of over-fitting. Given only 267 peer reviews in our case compared to more than ten thousand product reviews (Kim et al., 2006), this is an important consideration.

Though our absolute quantitative results are not directly comparable to the results of (Kim et al., 2006), we indirectly compared them by analyzing the utility of features in isolation and combined. While STR+UGR+MET is found as the best combination of generic features for both types of reviews, the best individual feature type is different (review unigrams work best for product reviews; structural features work best for peer reviews). More importantly, meta-data, which are found to significantly affect the perceived helpfulness of product reviews (Kim et al., 2006; Danescu-Niculescu-Mizil et al., 2009), have no predictive power (in isolation) for peer reviews. Perhaps because the paper grades and other helpfulness ratings are not visible to the reviewers, we have less of a social dimension for predicting the helpfulness of peer reviews. We also found that SVM regression does not favor ranking over predicting helpfulness as in (Kim et al., 2006).

In this study, we use expert-helpfulness ratings (the average of two expert-provided ratings) as our gold standard of peer-review helpfulness[10], though there are other types of helpfulness rating (e.g., author perceived helpfulness) that could be the gold standard as well. Our follow-up analysis on History2008 investigated the impact of different gold standards – expert-helpfulness ratings vs. student-helpfulness ratings – on the utility of different feature types for automatic review helpfulness prediction (Xiong and Litman, 2011b). In that analysis, we found that while simple linguistic features such as review length and the number of review sentences are the most predictive features when modeling students' perceived helpfulness; theory supported peer-review constructs are more useful in experts' models. With respect to related area of automated essay scoring (Attali and Burstein, 2006), others have suggested the need for the use of validated features related to meaningful dimensions of writing, rather than low-level (but easy to automate) features. In this perspective, our work poses challenge to the NLP community in terms of how to take into account the education-oriented dimensions of helpfulness when applying traditional NLP techniques of automatically predicating review helpfulness. These are interesting research topics that we would like to explore in our future work (will be further discussed in Chapter 6). In addition, we would like to emphasis that predictive features of perceived helpfulness are not guaranteed to capture the nature of "truly" helpful peer reviews (in contrast to the perceived ones).

3.6 SUMMARY

In this chapter, we demonstrate that techniques used in predicting product review helpfulness ranking can be effectively adapted to the domain of peer reviews, with minor modifications to the semantic and meta-data features. Our quantitative results shows that the generic features and our proposed peer-review specialized features are significantly

[10] Averaged ratings are considered more reliable since they are less noisy.

correlated with review helpfulness (in terms of both Pearson correlation and Spearman correlation). Our qualitative comparison between the product review and the peer review shows that the utility of generic features (e.g., meta-data features) in predicting review helpfulness varies between different review types. We further verify that prediction performance could be improved by incorporating specialized features that capture helpfulness information specific to peer reviews. In addition to the predictive power, these features are also theory-motivated, which better serves the educational purpose of the helpfulness model when used in online peer-review environment. Also, the proposed peer-review helpfulness model is low in dimensionality and thus suited for smaller corpora (compared to product reviews) that are typical in the peer review domain.

4.0 A GENERAL FEATURE REPRESENTATION OF REVIEW TEXTUAL CONTENT FOR REVIEW HELPFULNESS PREDICTION

In this chapter, we take a different path in predicting review helpfulness. Instead of developing auxiliary features that are specialized to a particular domain, we propose a general feature representation that can be obtained in the three examined domains, for predicting review helpfulness based on review textual content (referred to as content for the rest of the thesis). Specifically, the general content features characterize review language usage, content diversity and helpfulness-related topics comprehensively with respect to different content sources within the review. We examine the predictive power of the proposed features in comparison with reviews' superficial semantics across three domains.

Our experimental results show that the proposed features suit the prediction task better than the generic features from the previous chapter, especially in movie reviews and peer reviews. Our helpfulness-related topics show potential usefulness in assessing review helpfulness at the sentence level, which will be exploited in our new model of review summarization proposed in Chapter 5. Further, we observe that extracting features from different review content sources impacts review helpfulness prediction differently; differentiating content sources further increases our features' power for predicting review helpfulness significantly.

4.1 RELATED WORK

4.1.1 Content factors of review helpfulness

In this chapter, we take a closer look at various kind of computational linguistic features used in prior work for review helpfulness prediction. As we discussed in the previous chapter, unigrams are found to be quite predictive of review helpfulness for product reviews (Kim et al. 2006); while other syntactic and semantic features (sentiment words and domain lexicons) features are also predictive, using them in combination with review unigrams decrease the performance achieved by using unigrams alone. However, our experiments on peer reviews suggest that *high-level* representation of review content, such as lexical categories, is preferred over *low-level* lexicon-based features (e.g., unigrams). Replacing review unigrams with the lexical category features improves the model when other types of features are also used. As the lexical categories proposed in Section 3.2.2 are specialized to educational peer reviews and require human editing, in this chapter we wonder if there is any form of fully-automated high-level feature representation generalizable across domains. Here we focus on review text only, considering review "content" as the meaning expressed in review text, and we consider "content features" as computational linguistic features derived from review text, except structural features (e.g., the structural features in Table 3.1).

A lot of later studies focused on developing lexical features from a subset of review lexicons to capture a review's relevance to the review subject or other reviews using a bag-of-words approach (Liu et al., 2007; Zhang, 2008; Tsur and Rappoport, 2009; Zeng and Wu, 2013), assuming that a good review should have more information about the subject and use the exact terminology. However, such models are still comparatively high in feature space dimensionality, and only exploit reviews' superficial semantics.

Other properties of review content (in addition to review's relevance) were also studied for predicting review helpfulness. Zhang (2008) and Ghose and Ipeirotis (2011) investigated the subjectivity of reviews, which was found to have significant influence on review helpfulness. Several studies on IMDB movie reviews investigated reviewers' writing style, which was modeled by shallow syntactic features based on part-of-speech tags (Liu et al., 2008; Yu et al., 2012). Recently, Zeng and Wu (2013) identified that language and writing styles are two frequently mentioned reasons for some reviews being perceived as helpful in product interviews. They (Zeng and Wu, 2013) also mined the comparison style in reviews using regular expressions. In addition, review readability features, such as spelling errors, language formality, etc. were examined (O'Mahony and Smyth, 2010; Ghose and Ipeirotis, 2011) as well. However, the utility of these content features (Zhang, 2008; Ghose and Ipeirotis, 2011; Liu et al., 2008; Yu et al., 2012; Zeng and Wu, 2013) mentioned above were never compared to a simple lexical baseline.

Considering the dominance of review unigrams in prior helpfulness research (Kim et al., 2006) and in other content-based tasks (Louis and Nenkova, 2013), our experiments explicitly examine whether property-inspired content features (by themselves or in combination) can outperform unigram-based features in modeling helpfulness. However, as different works used different features for different types of reviews and the helpfulness prediction tasks were set up in various ways, it is not clear which helpfulness model is state-of-the-art that we should refer to. Therefore, we propose our own content features that can be generalized across domains.

Danescu-Niculescu-Mizil et al. (2009) pointed out that review content is not the only explanatory factor to reviews' perceived helpfulness. The review's timeliness (Liu et al., 2008), reviewer expertise (Liu et al., 2008; Ghose and Ipeirotis, 2011) and identity, the social network of reviewers and reviews (Lu et al., 2010) and the relation of the helpfulness rating to other ratings (Kim et al., 2006; Danescu-Niculescu-Mizil et al., 2009) also matter. However, such non-textual information is beyond the scope of our study. Because such metadata is not available for all types of reviews, and also in this thesis we investigate automatic review helpfulness prediction from the NLP perspective.

4.1.2 Content analysis in general

4.1.2.1 Lexicon dictionaries In terms of modeling textual content in general, our work is inspired by other text mining tasks both inside and outside the NLP community. One common approach is to utilize manually crafted dictionaries. For example, General Inquirer Word Counts[1] and MPQA[2] are used in sentiment analysis (Alm et al., 2005; Wilson et al., 2005). LIWC (Pennebaker et al., 2001) clusters terms based on syntactic and semantic functions in language, which is widely used for interpreting text data in disciplines such as Psychology, Social Science, etc. Because LIWC covers both affective processes (LIWC identifies positive and negative emotional words, which support sentiment analysis as well) and cognitive processes (similar to the keywords that we constructed for educational peer reviews), we choose LIWC in our presented work to characterize the general language usage in reviews.

4.1.2.2 Topic modeling Another popular approach to analyze text is through statistical topic modeling. Early work (Lin and Hovy, 2000) focused on identify the topic terms (known as topic signatures) of a target corpus that distinguish the corpus from general corpora. It is assumed that there is one dominant topic in the target corpus; all terms in the target corpus are either topic-relevant or topic-irrelevant. Under such single-topic assumption, the topic signatures can be identified by comparing the word distribution in the target corpus against a external background corpus. Later studies considered a document in terms of multiple hidden topics, and proposed various graphical models to infer the topics directly from the corpus of interest. The later approach is widely used for review analysis: while pLSA was used to predict the number of votes received by reviews (Cao et al., 2011), many Bayesian LDA-based models were proposed for sentiment analysis and opinion summarization (Mei et al., 2007; Lu and Zhai, 2008; Titov and McDonald, 2008a,b;

[1] http://www.wjh.harvard.edu/inquirer/homecat.html

[2] http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/

Blei and McAuliffe, 2010; Brody and Elhadad, 2010; Mukherjee and Liu, 2012; Sauper and Barzilay, 2013). Blei and McAuliffe (2010) proposed supervised LDA (sLDA) which models a review's sentiment score as a linear combination of the review's topics. In general, sLDA introduces document annotation as supervision in LDA's topic inference. By conditioning the topic sampling of each word on its document's annotation, the model is able to learn topics predictive of the annotations gradually. A lot of other statistical models were also proposed to infer review topics while predicting review sentiment in various granularities (Titov and McDonald, 2008b).

With respect to our work, we use the first approach (Lin and Hovy, 2000) (under the single-topic assumption) to identify external-content keywords from external materials; we use the second approach (under the multi-topic assumption) to identify review hidden topics from the review corpus itself. In particular, for the latter one, we choose supervised LDA (Blei and McAuliffe, 2010) as an initial attempt to combine topic modeling and review helpfulness analysis together.

4.1.2.3 Content categorization One popular categorization of user generated content is based on subjectivity. Ghose and Ipeirotis (2011) observed that review subjectivity has different utility in helpfulness prediction for different types of reviews: objective content is preferred for "feature-based subjects" (e.g., electronics) while subjective content is preferred for "experience subjects" (e.g., movies). We suspect that the preference for subjective content in the reviews of "experience subjects" might be attributed to reviewers' indirect quotations (or descriptions) of the content of review subjects (e.g., paper content for peer review), which tend to be subjective (known as expressive subjective elements (Wiebe et al., 2005)). Therefore, we propose to categorize review content regarding whether it refers to the review subject or the review-subject's content. While early work in opinion mining and sentiment analysis extracts and distinguishes opinion expressions with respect to the opinion holder (known as opinion sources) (Wiebe et al., 2005; Choi et al., 2005; Bethard et al., 2004), in this work, we consider the full review content (both opinions and factual statements) rather than opinions alone, to examine if the proposed content categorization 4.5 would make an impact on review helpfulness prediction.

4.2 DATA

This chapter considers three distinct review domains: Amazon product reviews, IMDB movie reviews and educational peer reviews from the History class (History2008). Descriptive statistics of our three data sets are provided in Table 2.1, Chapter 2. As it shows, the size of the camera reviews is much larger than the other two, and the movie reviews are longest among the three in general. In addition, camera reviews (average=.80, std=.28) and movie reviews (average=.74, std=.16) tend to receive higher helpfulness ratings than peer reviews (History2008) (average=.43, std=.24).

4.3 FEATURES

In total, we develop 104 computational linguistic features to create a compact representation of review content in contrast with unigrams (more than two thousand) for modeling review helpfulness. In this section, we explain our motivation and how we formalize the features.

4.3.1 Representing review language usage (LU)

We suspect that the predictive power of review unigrams might be attributed to only a subset of lexicons. In the previous chapter, we observed that the unigram features and the lexical category features are of similar predictive power, while replacing review unigrams with the lexical categories improves the predictive performance when other kinds of features were also used (as shown in Table 3.4). However, the lexical categories are especially designed for educational peer reviews and they still require manual editing. To fully automatically categorize lexicons based on their syntactic and semantic functionality in general, we refer to the 82 language dimensions in LIWC, which is a publicly available dictionary (developed by linguistic experts) widely used for text analysis in psychology (e.g., examining one's mental states (Tausczik and Pennebaker, 2010)), cognitive science (e.g., detecting lies in police interviews (Vrij et al., 2007)), social science (e.g., analyzing marital interactions (Simmons et al., 2005)), education (e.g., assessing cohesion in writing (Graesser et al., 2004)), etc.

LIWC categories cover both linguistic processes and psychological processes. With respect to linguistic processes, LIWC not only counts dictionary words ("Dic"), fillers and words greater than 6 letters ("Sixltr") – which capture language standardness and complexity, but also categorizes function words (e.g., personal pronouns, articles, past tense, adverbs, negations, etc.) and punctuation – which reflect writers' personal states (Tausczik and Pennebaker, 2010) and have been used to characterize reviews' writing style (Otterbacher, 2010). With respect to psychological processes, the affective processes ("affect") look at semantic subjectivity ("posemo", "negemo") and affect (e.g., "anger"), while cognitive processes ("cogmech") – based on mining words such as *know*, *ought*, *because*, *should*, *maybe*, etc.– are considered useful for review helpfulness analysis (Ando and Ishizaki, 2012).

For our modeling of review helpfulness, we compute the 82 LIWC counts for each review d and normalize the LIWC counts by the total word count as one type of our content features.

$$f_{LU(c)}(d) = \frac{\sum_{w \in LIWC(c)} count(w)}{\sum_{w \in d} 1} \tag{4.1}$$

As an example, consider the two movie reviews ($review_1$ denotes the more helpful one; $review_2$ denotes the less helpful one) provided in Section 2.2 with respect to LIWC category "percept" (perception, including *observing*, *heard*, *feeling*, etc.): $f_{LU(percept)}(review_1) =$

Domain	**LIWC categories**
Camera	AllPct Apostro Dic Numerals Period Sixltr WC achieve affect anger auxverb bio certain cogmech excl ...
Movie	hear humans QMark anger ingest death relig percept.
Peer	AllPct Dic Period SemiC Sixltr WC affect cogmech funct insight past posemo relativ social verb ...

Table 4.1: LIWC categories with significant Pearson correlation (r) and Spearman correlation with review helpfulness ratings ($p \leq .05$), in descending order of r.

.003, $f_{LU(percept)}(review_2) = .111$. $Review_1$ contains a smaller percentage of perception words than $review_2$, which suggests that using more perception words may have a negative impact on reviews' perceived helpfulness.

In total we observe that LU correlates significantly with helpfulness for 53 LIWC categories for camera reviews, but only with 8 for movie and 21 for peer reviews (top row of Table 4.10). This suggests that movie and peer reviews are more difficult compared to the camera reviews, as word usage alone is not enough to explain review helpfulness. Table 4.1 provides some examples of the significant LIWC categories.

4.3.2 Representing review content diversity (CD)

Studies (Carenini et al., 2006; Zeng and Wu, 2013) show that people prefer reviews that cover multiple aspects and provide enough detail. The magnitude of content diversity has also been used as a criteria for selecting useful review elements in review summarization (Carenini et al., 2006; Lerman et al., 2009), as well as for characterizing helpful educational peer reviews (Ramachandran et al., 2013). Therefore, we expect that content diversity is predictive of review helpfulness ratings. For each review d, we compute language entropy over word distribution to reflect its content diversity (Formula 4.2). Such a lexical statistic was used for measuring review content richness (Otterbacher, 2010) and shown to be effective in measuring the content variety of telephone conversations regarding different social relationships between speakers (Stark et al., 2012) as well. We expect it to be also useful in our analysis of review helpfulness. We use $f_{CD}(d)$ and its normalized value (by review word count) as our second type of content feature.[3]

$$f_{CD}(d) = -\sum_{w \in d} p(w|d) \log p(w|d) \tag{4.2}$$

Considering the same two movie reviews in Section 2.2, the extracted corresponding content diversity features are (7.5, .01) from the more helpfulness one, and (5.6, .09) from the less helpfulness one, in which the first number is the absolute value and the second is the normalized value.

4.3.3 Mining helpfulness-related review topics (hRT)

To discover review topics' discrimination of review helpfulness, we introduce supervised LDA (Blei and McAuliffe, 2010) to learn review topics that are associated with review helpfulness ratings. As sLDA is shown to be effective in learning review sentiment at the topic level (Blei and McAuliffe, 2010), we are curious about whether this technique can be applied on review helpfulness as well.

To extract the helpfulness-related topics, we train sLDA on a training set with the helpfulness gold-standard as the document annotation for supervision, using 20 topics (t_k, K=20) and the best hyper-parameters that we learned in a pilot study[4]. Then we use the

[3] Because there is no significant difference in the predictive power between the two statistics while they do vary with reviews in different ways, we consider both of them to represent content diversity in this work.

[4] We implemented sLDA using the topic modeling framework of Mallet (McCallum, 2002), and set the parameters based on our pilot study of LDA on the same data sets. We set the topic-specific priors to 0.5 and word-specific distribution priors are set to 0.1. The inference is run for 100 iterations in both the Estimation and the Maximization steps. We also experimented with asymmetric topic priors which were dynamically optimized through training, however, the resulting topics are less predictive compared to using symmetric topic priors.

learned model (M) to infer the topics on the test set without the document annotations (same as LDA). In our helpfulness model, we use the inferred posterior topic distribution in each review d (Formula 4.3) as the third type of our content features.

$$f_{hRT(k)}(d) = \sum_{w \in d} p(z = t_k | w, M) \tag{4.3}$$

Figure 4.2 shows a 20-topic sLDA model fit to our movie reviews: topics are presented as their 10 most likely words (on the y-axis), and are associated with their estimated coefficients (η_k) in the linear regression of sLDA (on the x-axis). Figure 4.2 shows that some topics are about movie plots, though their topic words also include evaluation terms, indicating the heterogeneous nature of movie review content. Similar examples are provided for camera reviews (Figure 4.1) and peer reviews (Figure 4.3). When comparing across domains, we notice that camera reviews have more topics predictive of review helpfulness (19 out of 20 coefficients are in the range between 0.7 and 1.0), while both movie reviews and peer reviews contain quite a few topics that are of little predictive power (more than 9 coefficients are smaller than 0.2).

4.3.3.1 Inferring sentence-level review helpfulness Note that the topics are learned in the supervision of *review-level* helpfulness ratings. Nevertheless, they are useful in differentiating review *sentence-level* helpfulness as well, as suggested by examples from Tables 4.2 to 4.4, one for each review domain. For each review sentence s, we estimate its helpfulness score by applying the linear regression model (learned by sLDA) on its inferred topic assignments (Formula 4.4). While the review example in Table 4.4 is considered one of the most useful reviews (helpfulness rating = 1), the second sentence (bolded) is predicted as the most helpful one. Due to the lack of sentence-level helpfulness gold-standard, we pursued an extrinsic evaluation of sentence-level review helpfulness prediction in review summarization tasks (Chapter 5).

$$H(s) = \sum_{k} \eta_k \sum_{w \in s} p(z = t_k | w, M) \tag{4.4}$$

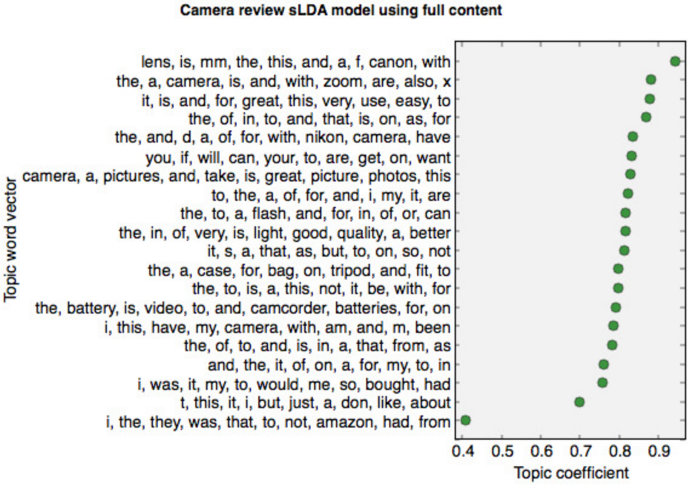


Figure 4.1: Topics and coefficients learned from Amazon camera movie review data.

4.3.3.2 sLDA analysis Wallach et al. (2009) suggested using asymmetric priors (α) for document-specific topic distributions in LDA, as in practice some topics are likely to occur more frequently than the others. We wonder if this also applies to the supervised topic modeling settings. As the original sLDA proposed by Blei and McAuliffe (2010) (denoted as sLDA-sym) uses symmetric α, we conduct additional analysis to see if incorporating asymmetric α_i into sLDA and dynamically optimizing them during training can yield a better model.[5]

However, our 10-fold cross-validation results based on per-word log likelihood and predictive R^2 scores (Table 4.5) show that it seems to be a trade-off between a fit model and good prediction of review helpfulness. While sLDA-asym fits the data better (higher per-word log likelihood ($p < .05$)) on camera reviews and peer reviews, its topics have little correlation with review helpfulness. It seems that the asymmetric setting of the topic priors

[5] We use the optimization procedure provided by Mallet with its default parameter settings. For sLDA, α_i is initialized equally for each dimension.

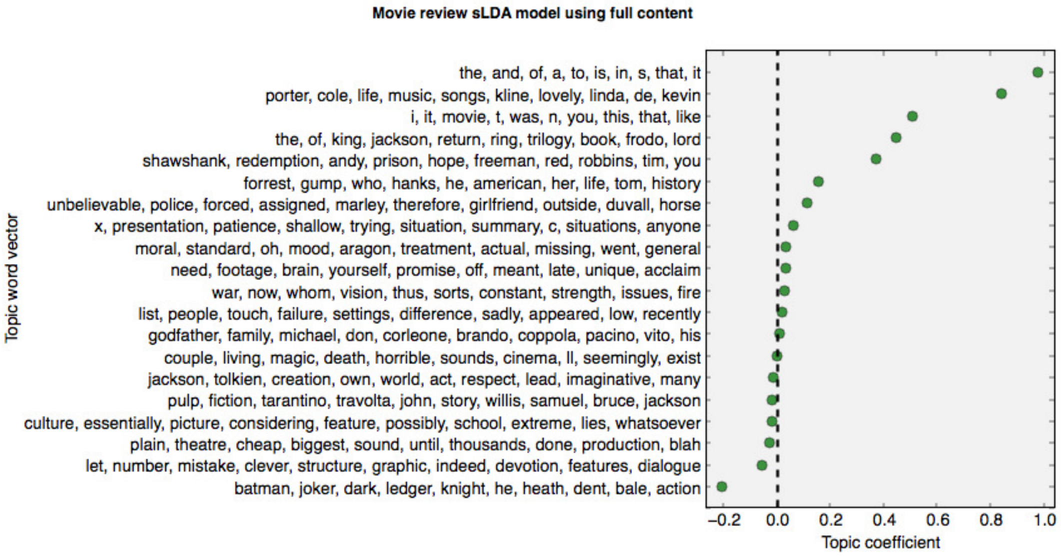


Figure 4.2: Topics and coefficients learned from our IMDB movie review data.

conflicts with the supervision goal of sLDA. Our analysis results suggest that asymmetric topic priors do not help train supervised topic models for prediction tasks.

4.4 HELPFULNESS PREDICTION EXPERIMENTS

In this chapter, we consider similar machine learning experiments as we used in the previous chapter.

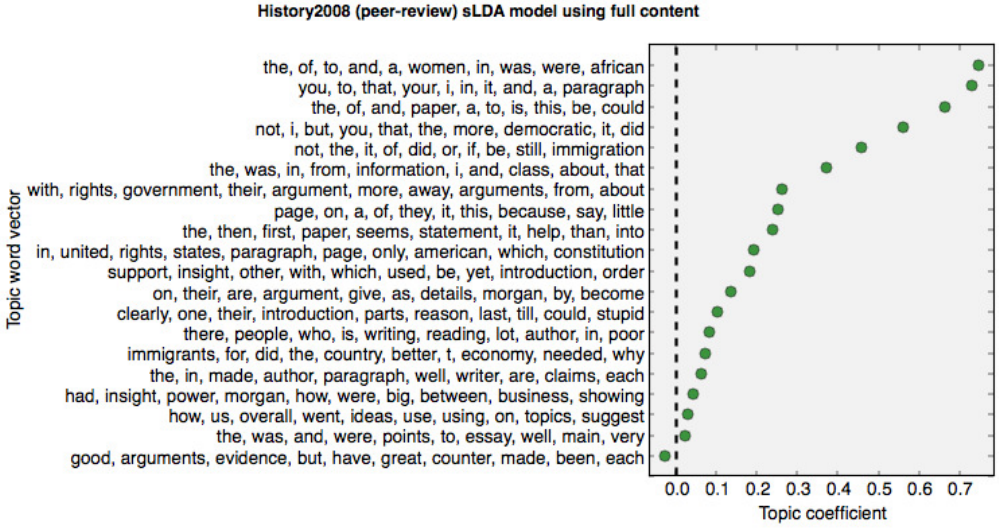


Figure 4.3: Topics and coefficients learned from History2008 peer-review data.

4.4.1 Experimental setup

With respect to our generalization approach for review helpfulness prediction, we hypothesize that:

H3 Review helpfulness can be predicted using only review text, based on the same computational linguistic representation across domains.

H4 The proposed content features outperform review unigrams.

Since Kim et al. (2006) and we (Xiong and Litman, 2011a) used SVM regression with a radial basis function kernel provided by SVM^{light} to train a helpfulness model based on various type of features for camera and peer reviews, respectively, we use the same setting for our machine learning experiments.[6] For evaluation, we use 10-fold cross validation.

[6] All features are transformed logarithmically and normalized from 0 to 1 as in prior work (Kim et al., 2006).

Considering that training sLDA models is quite time consuming, in this chapter, we use the same 10-fold split for the cross-validation in all experiments.

To test Hypothesis **H3**, we experiment on not only the educational peer reviews, but also camera reviews and movie reviews (both referred to as customer reviews). To verify Hypothesis **H4**, we compare the proposed content features (LU, CD and hRT) against two baseline feature sets from (Kim et al., 2006) as introduced in the previous chapter. Note that these features could also be generalized to all three of our domains.

We report the results based on Pearson correlation (r), both mean and standard deviation (within parenthesis). As we observe similar results when using Spearman correlation, we omit reporting it here.

4.4.2 Results

First we present our results compared to the unigram baseline in Table 4.6. Recall that Kim et al. (2006) found that unigrams would suppress other text-based features (e.g., syntactic and semantic) even when included. Here we evaluate each type of our proposed content features in isolation, as well as in combination (denoted as *content*). All of the results in the first four rows (Table 4.6) significantly correlate ($p < .05$) with review helpfulness ratings except CD on movie reviews. These results verify our hypothesis about the generality of the proposed overall feature representation for predicting review helpfulness across domain (Hypothesis **H3**).

When compared across domains, however, the utility of these features is not the same (Hypothesis **H3** is only partially supported): LU and CD are most predictive on peer reviews and least predictive on movie reviews, while hRT work better on movie and peer reviews than camera reviews. Such differences suggest that review topics play a more important role in the helpfulness of the former two domains than they do in camera reviews. However, it can also be explained by the heterogenous nature of the review content (Section 4.5) which makes movie/peer reviews difficult to analyze.

When the significant content features are compared against the unigrams, they outperform the unigrams on movie reviews and peer reviews ($p < .05$), but the unigrams still work best on camera reviews. It seems that the superficial semantics are good enough for capturing the review helpfulness of camera reviews, which is also the case in Kim et al.'s analysis on their own Amazon product review corpora (MP3 players and digital cameras). Also, considering the data size, using unigrams on our movie/peer reviews might have caused over-fitting. However, when we apply downsampling on the camera reviews from 4050 to 280 reviews, the unigrams still work best ($r = .69$, $p = .001$), while the correlations between our content features and the helpfulness ratings are no longer significant ($p < .05$).

As a stronger baseline, we further compare our content features versus the unigrams by combining each with structural (STR), syntactic (SYN), semantic (domain (DW) and sentiment (SENT) lexicons) and meta (MET) features explored in (Kim et al., 2006). As Table 4.7 shows, the pattern of Table 4.6 still holds.

Therefore, we conclude that Hypothesis **H4** is only partially supported: our content features outperform the unigram features, with/without the other types of generic features, though only in the movie and the peer-review domains. Review unigrams are still the most predictive features for predicting product review helpfulness.

4.5 DIFFERENTIATING REVIEW'S INTERNAL CONTENT AND EXTERNAL CONTENT

Inspired by related work discussed in Section 1.2.2, we propose to analyze review content regarding whether it is the reviewer's evaluation of the review subject (e.g., *"This is the best camera I've ever had!"*) or it merely refers to the review subject and its content (e.g., movie plot) as evaluation context for review helpfulness prediction (e.g., *"Schultz tells Django to pick out whatever he likes."*). Specifically, we differentiate these two kinds of content during feature extraction, and the corresponding feature sets are examined in isolation and in combined for predicting review helpfulness. In this section, we focus on the two domains (movie and peer) that benefit from our new features identified earlier to demonstrate the merit of this approach in the task of review helpfulness prediction, though we still include the experimental result on camera reviews for completeness.

4.5.1 Internal content vs. external content

While early work in opinion analysis extract and distinguish opinion expressions with respect to the opinion holder (known as *opinion sources*) (Wiebe et al., 2005; Choi et al., 2005; Bethard et al., 2004), in this work we consider the full review content (both opinions and factual statements) rather than opinions alone. Therefore we denote the two kinds of content that we aim to differentiate in this work as "internal content" and "external content".

While the internal content of a review is the reviewer's personal experience or evaluations of the review subject, the external content is the reviewer's references (or paraphrases) of the review subject. With respect to the terminology used in opinion analysis (Wiebe et al., 2005), the internal content contains the reviewer's objective speech events (such as "*I'm merely a birthday - holiday type picture taker*") and direct subjective expressions (e.g., *"This is the best camera I've ever had!"*). As the external content refers to the review subject (e.g., review aspects) or directly/indirectly quotes of the subject content (e.g., movie plot), it could contain objective speech events (e.g., *"Schultz tells Django to pick out whatever he likes."*) and subjective expressions (e.g., "... the main point was that *the enslavement of African Americans, the fight for women's suffrage and the immigration laws that were passed greatly effected the U.S. democratically.*") from the review subject, as well as expressive subjective elements that reveal the reviewer's opinion towards the subject ("We learn about the *true bravery and potential of* hobbits as Merry helps cut down the Witch King"). Despite of the mixture of different opinion sources in the external content, we argue that the external content generally serves as the context for reviewers to express their explicit opinions about review subjects. When a review subject also has content (e.g., books, movies, essays, etc), the review external content could be versatile. As the internal content and the external content might play different roles in review sense making, such heterogeneous review content poses challenges to review analysis (Turney, 2002; Pang and Lee, 2004).

4.5.2 Identifying review external content

Completely splitting the internal and external content could be hard, which might involve fine-grained opinion analysis. As an initial investigation of its impact on the helpfulness prediction task, we consider reviews' external content in terms of *keywords* (subject/domain related terms) that can be automatically extracted from external resources of the review subject. Given a set of the external-content keywords, we identify their occurrences in each review as the review's external content, and consider the remaining words in the review as its internal content. Therefore, within a given review, the internal content and external content are exclusive (at the word-level). In this work, we reduce the problem into keyword extraction from the external content. Although other approximation methods can be used, we leave such exploration for our future research.

As the subjects of peer reviews are papers from the same assignment, we extract the keywords from the papers as a whole. We exploit a corpus-based topic signature extraction algorithm (Lin and Hovy, 2000), using all student essays as the target corpus, and 5000 documents from the English Giga-word Corpus as the background corpus. For movie reviews, we take advantage of the keywords and plot text (summaries and synopses) available on IMDB.com, and create a keyword set for each movie using their keywords and the actor/actress names highlighted in the plot text. Plus, we augment each keyword set with the plot topic signatures extracted from the plot text and the related reviews, using the same extraction algorithm with all movie reviews as the target corpus (all other settings are the same as the peer reviews). For camera reviews, although we argue that the review content is less heterogenous as most of camera review content is internal, we still extract the product keywords and treat them as external content for a complete comparison across all three domains, given the the particular approximation used here. We first apply the same topic signature extraction algorithm with all the camera reviews to be the target corpus (all other settings are the same as the peer reviews), and then we eliminate sentiment terms that are categorized as "Positive" or "Negative" in General Inquirer Dictionaries from the extracted topic signatures. The vocabulary size of the internal and the external content for each domain is summarized in Table 4.8. Note that for peer reviews and camera reviews, the external keywords are extracted for the whole corpus, while for movie reviews, the external keywords are extracted for each movie separately. Therefore, in the former two domains, the vocabulary size of the internal content and the external content add up to the vocabulary size of the full corpus in Table 1, but this is not the case in the movie review domain. Examples of the external content lexicons are shown in Table 4.9.

To illustrate the impact of the proposed content categorization, in Table 4.10 we show the difference between review internal and external content in the total number of significant LIWC categories (discussed in Section 4.1) that they yield. Table 4.10 shows that more significant categories are observed for peer reviews when applying LU analysis on reviews' internal content only, compared to using the full content without such differentiation. Also for movie reviews, the external has most significant categories.

4.6 EVALUATION ON DIFFERENT CONTENT TYPES

In this experiment, we examine the impact of review content differentiation (internal content vs. external content) when extracting the proposed content features (Section 4.3) for predicting review helpfulness across the three review domains.

4.6.1 Experimental setup

H5 Distinguishing review-subject descriptions and other review content facilitates review helpfulness prediction.

To test the hypothesis above, we extract our content features (LU, CD and hRT) in four different ways: 1) from the full content of all reviews (F), 2) from the internal content only (I), 3) from the external content only (E) and 4) from both the internal and the external content but separately (I+E). We compare the utility of the two types of content based on their corresponding features' predictive power of review helpfulness in the same SVM regression setting as we used in Section 4.4. Note that the feature vectors generated from I+E are equivalent to concatenating the feature vectors generated from I and E.

4.6.2 Results

Experimental results are provided in Table. 4.11 Comparing the internal content (I) and the external content (E), we always get much more predictive features from I than E in peer reviews, while it seems to be the opposite in movie reviews when examining each feature type separately. (No general pattern is observed in camera reviews.) In particular, I yields the most predictive topics (hRT) for peer reviews but the least predictive ones for camera reviews and movie reviews (customer reviews). Considering the educational context of the peer review, this suggests that while what external content is mentioned is important to review helpfulness ($r = .28$, $p \leq .001$), how the external topic content is discussed (the internal content) is more crucial ($r = .53$, $p \leq .001$).

More importantly, for both movie reviews and peer reviews, differentiating reviews' internal content and external content yields improvement on all features and their combination except CD. In spite of the proposed feature types' individual differences across domains, their combination (content) always achieves the best performance when these features are extracted from both the internal and the external content, separately (I+E).

To conclude, the machine learning experiment results confirm that different review content types have different predictive power of review helpfulness, which varies with the review domain as well. We conclude that content differentiation is an important procedure to take before feature extraction for building review helpfulness models, which improves the prediction performance for camera reviews, movie reviews as well as educational peer reviews.

4.7 DISCUSSION

In contrast with our specialization approach on the same educational data set in the previous chapter, although the experimental setup is slightly different (different setups for cross-validation)[7], we can still compare our content-based general model with the peer-review specialized model (row 7 in Table 3.4) for predicting peer review helpfulness by considering "STR+UGR+MET" (row 4 in Table 3.4 and the first row in Table 4.7) as the anchor. While our specialized model improves the prediction performance from the baseline by 8%, the general model outperforms the baseline by 7% (in Table 4.7) and by 10% (in Table 4.11) in terms of relative improvement, after differentiating the internal and external content.

With respect to feature engineering, LIWC (Table 4.1) can be viewed as a generalized/standardized version of the peer-review specialized lexical categories (Table 3.2). Certain categories of LIWC (e.g., affective processes) and the helpfulness-related review topics replace the lexical semantic features (e.g., TOP in Table 3.4) used in the previous approach. Although the cognitive-science features are not directly engineered in the general model, certain aspects are represented in terms of the cognitive-process related LIWC

[7] In Section 3.3, we randomly split the data set into 10 folds for each individual trial (e.g., each row in Table 3.3). However, in this chapter we use the same 10 folds in all trials across experiments, because it takes a long time to train a sLDA model.

categories, as well as some of the latent topics learned by supervised LDA. For instance, in Figure 4.3, “page”, “paragraph”, and domain-specific terms were picked up in the 10th topic (from the top of the y-axis) indicating problem localization; “good”, “great” and “argument” in the last topic suggest praises on argumentation. Furthermore, it is important to note that the content features presented in this chapter is fully automated, while the peer-review specialized features require peer-review domain knowledge and need human editing. However, in terms of the computational cost, the content features do rely on sophisticated natural language processing techniques which is computationally more expensive in terms of both time and space.

Considering the pros and cons between the two approaches, if instant prediction is required, the specialization approach would serve the needs better, with a little sacrifice of the performance. Also, for certain domains that have clear definition of review helpfulness and in which the review topics are constrained (e.g., reviews under specific instruction), the first approach could suit the prediction task better. In the opposite, when the definition of review helpfulness is obscure and reviewers have more freedom in what they can write, the generalization approach – purely data-driven – would serve the needs better.

Since the movie review data used in our work is comparatively small (to be comparable to the size of the peer reviews), in the future we would like to run experiments in larger scale to see if the data size matters. Though when we down-sampled the camera reviews, unigrams still performed best. While we do not intrinsically evaluate our supervised topic model for inferring review helpfulness at the sentence level, we will use the sentence-level helpfulness scores for content selection in an extractive summarization system in the next chapter, which serves as extrinsic evaluation.

4.8 SUMMARY

In this chapter, we proposed a new review helpfulness model using features extracted from review content only, by characterizing review language usage, content diversity and helpfulness-related topical information. We showed that the three new content representations work well in multiple review domains, and better than unigrams (both when compared directly, or when used in conjunction with prior content as well as meta features) for domains that involve more heterogeneous review content (e.g., movie reviews and peer reviews). In addition, we proposed to extract the content features from different content types separately, which are categorized based on whether the review content is only referring to the review subject. We showed that applying the proposed content categorization before feature engineering yields significant ($p \leq .05$) improvement in the helpfulness prediction task.

No.	**Sentence text**	$H(s)$
1	**I have another camera w/12x optical & tons of features but I wanted something compact & ready to take quick snaps.**	**.82**
2	This fits the bill.	.13
3	On my recent vacation i was impressed with how quickly it booted up and focused in to get those unplanned shots.	.37
4	The x-large screen makes it really easy to see if you got a good one and the image stabilisation seems to work better than on my previous camera (canon A80) ... either that or my hand is steadier.	.38
5	Last but not least - the battery life was impeccable.	.07
6	I took a nearly 200 photos, plus spent a lot of time reviewing and showing off pics to friends with no need to recharge.	.17
7	The battery is a ‘custom’ one which concerned me but the charger is VERY compact and travels well.	.11
8	It has integrated prongs that fold flat when not in use - no cables.	.07

Table 4.2: Estimating sentence-level helpfulness scores using sLDA trained with review-level helpfulness ratings. Sentences are segmented from a Camera review example with helpfulness rating = 1.

No.	Sentence text	$H(s)$
1	**He may have been good in Philadelphia but be is excellent in Forrest Gump.**	**.66**
2	Tom Hanks delivers another great performance in his career by portraying the lovable , king yet not so intelligent character Forrest Gump.	.37
3	It is also Tom Hanks' second straight win for the Best Actor Oscar which he becomes the second man to do said accomplishment after Spencer Tracy.	.27
4	Whilst not as dramatic as Philadelphia , Tom Hanks' performance is just as great in this movie and this movie could possibly be the film of Tom Hanks' career as he used to be a comedy guy who turned to drama in a way which would paved for future stars such as Jim Carrey (The Truman Show), Reese Witherspoon (Walk the Line) and Will Ferrell (Stranger Than Fiction) to name a few.	.34
5	Also staring in this great movie classic are Robin Wright-Penn who plays Jenny , Gary Sinise who was nominated for an Academy Award for his portrayal of Lieutenant Dan Taylor , Mykelti Williamson as Forrest's best friend and shrimper Benjamin Buford "Bubba" Blue and Sally Field as Mrs. Gump.	.16
6	This film was nominated for a total of thirteen Academy Awards but won six of them which include Best Film Editing , Best Visual Effects , Best Adapted Screenplay , Best Picture , Best Director-Robert Zemeckis and Best Actor-Tom Hanks.	.10
7	This is one masterpiece of a movie that will not be forgotten about in a long time.	.05
8	Bravo!	.003

Table 4.3: Movie review example of estimating sentence-level helpfulness scores. The review's helpfulness rating = .8.

No.	**Sentence text**	$H(s)$
1	There does not seem to be much logic behind the arguments being made.	.67
2	**The thesis should involve stating whether the United States was more, less, or equally democratic between 1865 and 1924.**	**.77**
3	Or at least, I am assuming that is the essay prompt you intended to chose.	.50
4	The paper is an excellent essay on immigration restriction, but it does not deal with the true issue at hand.	.51
5	The paper talks more about how the immigrants were misrepresented than how they were denied true democratic rights.	.60
6	More emphasis should be placed on the inequalities that immigrants experienced in voting and constitutional freedoms, otherwise the paper is completely off prompt.	.64

Table 4.4: Peer review (History2008) example of estimating sentence-level helpfulness scores. The review's helpfulness rating = 1.

	Model	**Per-word log-likelihood**	**predictive R^2**
Camera	sLDA-sym	-7.76(.006)	.124(.022)*
	sLDA-asym	-6.78(.028)*	.030(.013)
Movie	sLDA-sym	-7.25(.061)	.120(.143)
	sLDA-asym	-7.01(.016)	.145(.159)
Peer	sLDA-sym	-7.19(.064)	.244(.125)*
	sLDA-asym	-6.53(.021)*	.027(.027)

Table 4.5: Per-word log-likelihood and predictive R^2 of the review data sets. Reported values are the average and standard deviation (inside parenthesis) of scores from 10 cross-validation. Significantly better results between the two models for each domain and metric ($p < .05$) are highlighted with star.

Pearson r			
Feature set	**Camera**	**Movie**	**Peer**
LU	.469(.089)-	.197(.417)-	.599(.274)+
CD	.418(.087)-	-.033(0.451)-	**.612(.239)+**
hRT	.351(.082)-	.440(.305)+	.523(.241)
content	.490(.068)-	**.444(.394)+**	.599(.273)+
unigram	**.620(.043)**	.218(.553)	.518(.266)

Table 4.6: SVM regression performance (Pearson Correlation r) using the proposed content features. Reported values are the average and standard derivation (inside parenthesis) of scores from 10-fold cross validation. For each domain, the best feature set is highlighted in bold. Comparing with the unigrams, significantly better results are labeled with "+" and significantly worse results are labeled with "-".

Features	**Camera**	**Movie**	**Peer**
unigram+STR+MET (baseline)	**.635(.085)**	.234(.557)	.584(.231)
content+STR+MET	.574(.089)	**.470(.391)**	**.626(.231)**
unigram+STR+MET+SYN+DW+SENT	**.656(.081)**	.202(.548)	.550(.282)
content+STR+MET+SYN+DW+SENT	.615(.086)	**.435(.423)**	**.630(.242)**

Table 4.7: SVM regression performance (Pearson Correlation r) using all features. We use the best feature set reported in Kim et al. (Kim et al., 2006) for product reviews as our baseline.

Domain	**Internal content**	**External content**
Camera	9009	4151
Movie	8747	1659
Peer	2180	519

Table 4.8: Vocabulary size of reviews' internal content vs. external content.

Domain	**External keywords**
Camera	camera, lens, pictures, canon, mm, digital, battery, flash, zoom, price, video, image, , ...
Movie	merry, goondor, treebeard, helm, gandalf, wormtongue, allies, fangorn, gfrodo, war, ...
Peer	war, african, americans, women, democracy, rights, states, vote, united, amendment, ...

Table 4.9: Example of the external content (key)words.

Content type	**Camera**	**Movie**	**Peer**
full	**53**	8	21
internal content	51	10	**30**
external content	40	**16**	7

Table 4.10: Number of significant LIWC categories.

	Camera reviews			
Features	F	I	E	I+E
LU	.469(.089)	.476(.078)	.386(.121)	**.513(.088)**
CD	**.418(.087)**	.403(.095)	.406(.068)	.415(.076)
hRT	.351(.082)	.284(.125)	.314(.086)	**.354(.086)**
content	.490(.068)	.478(.080)	.465(.069)	**.516(.071)**

	Movie reviews			
Features	F	I	E	I+E
LU	.197(.417)	.301(.627)	**.414(.283)+**	.392(.412)+
CD	-.033(.451)	.047(.462)	**.115(.374)**	.094(.405)
hRT	.440(.305)	.418(.284)	.511(.280)	**.518(.268)+**
content	.444(.394)	.417(.397)	.253(.491)	**.523 (.311)+**

	Peer reviews			
Features	F	I	E	I+E
LU	.599(.274)	.620(.262)	.454(.141)-	**.632(.243)+**
CD	**.612(.239)**	.607(.220)	.284(.503)-	.586(.223)-
hRT	.523(.241)	**.529(.167)**	275(.381)-	.521(.193)
content	.599(.273)	.631(.255)	.447(.145)-	**.640(.251)+**

Table 4.11: Performance of features extracted from different content types. For each feature set, significant results ($p \leq .05$) compared with F are marked with "+" (better) or "-" (worse), and the best performance is highlighted in bold (F: full content, I: internal content, E: external content, I+E: internal and external content).

5.0 REVIEW SUMMARIZATION

As we believe that reviews' salient information can be found using their helpfulness ratings, in this chapter, we investigate two ways to introduce review helpfulness into a traditional multi-document extractive summarization framework: 1) use review-level helpfulness ratings to filter out unhelpful reviews before summarization, 2) use sentence-level helpfulness scores as features for sentence selection during summarization.[1] As shown in the previous chapter, using user-generated helpfulness assessment, sLDA can infer hidden topics that are predictive of review helpfulness, and our sentence-level helpfulness scores show potential in differentiating review helpfulness at the sentence level. Therefore, we expect the estimated sentence helpfulness scores to make good features in extractive review summarization for selecting the "helpful" text units. In contrast, for review-level helpfulness, we use user-provided helpfulness gold-standard rather then the predicted values in this chapter.

There are two main advantages of our helpfulness-guided approach: 1) it is user-centric, as we directly associate the information extraction process with user-generated feedback; 2) it is generalizable: while what is salient in reviews might differ from one domain to another, our supervision for content selection is merely the meta data of the reviews which is widely available[2] plus it requires little feature engineering.

To demonstrate our approach, we evaluate our hypothesis using MEAD – an open-

[1] The proposed method and evaluation results on customer reviews are published in COLING2014 (Xiong and Litman, 2014).

[2] If it is not available, we have shown that review helpfulness can be assessed fully automatically.

source multi-document extractive summarization framework based on which we implement the proposed ideas into two models respectively. We compare the helpfulness-guided summarizers against an advanced baseline provided by MEAD in both human evaluation and automated evaluation. Also, our work shows that the helpfulness-related topic words learned from the review-level supervision can capture review helpfulness at the sentence-level as well.

For the rest of this chapter, we will first compare and contrast our work with related work in the NLP literature, describe the experimental set-up, and then present our evaluation results on customer reviews (Camera reviews and Movie reviews) and educational peer reviews (Physics2014) separately.

5.1 RELATED WORK

Multi-document summarization is a classic NLP task aimed at extraction of salient information from multiple textual documents, which has been mostly studied for news articles. A key task is to identify important text units – content selection. Early extractive summarization techniques focus on identifying similarities between sentences, to identify common themes by clustering and then select the most representative sentence from each cluster (Radev et al., 2004). Later works use statistical models to identify the content structure (Barzilay and Lee, 2004) instead of clustering. Usually the novelty of a sentence to be selected is examined with respect to sentences that are already included, using maximal marginal relevance (Carbonell and Goldstein, 1998).

Prior successful extractive summarizers score a sentence based on n-grams within the sentence: by the word frequency (Nenkova and Vanderwende, 2005), bigram coverage (Gillick and Favre, 2009), topic signatures (Lin and Hovy, 2000) or latent topic distribution of the sentence (Haghighi and Vanderwende, 2009), which all aim to capture the "core" content of the text input. Other approaches regard the n-gram distribution difference (e.g., Kullback-Lieber (KL) divergence) between the input documents and the summary (Lin et al., 2006), or based on a graph-representation of the document content (Erkan and Radev, 2004; Leskovec13 et al., 2005), with an implicit goal to maximize the output representativeness. When the extraction idea directly applied to online reviews, Ando and Ishizaki (2012) manually annotated informative sentences in travel reviews to capture "what is salient" from user's point of view. In comparison, while our approach follows the same extractive summarization paradigm, it is metadata driven, identifying important text units through the guidance of user-provided review helpfulness assessment. Abstractive techniques have also been proposed for multi-document summarization. In addition to identifying salient text units from the input text, the abstractive methods further merge the text units by sentence editing and information fusion to make the summary more concise (Knight and Marcu, 2002). Because the focus of our research is to *select* useful review content by analyzing user-provided helpfulness assessment, we do not elaborate on the abstractive techniques in this thesis.

When it comes to online reviews, the desired characteristics of a review summary are different from traditional text genres (e.g., news articles), and could vary from one review domain to another. In general there are two major paradigms. One is by modifying existing multi-document summarization framework. Various methods have been proposed to generate review summaries of different desired properties, primarily based on opinion mining and sentiment analysis (Carenini et al., 2006; Lerman et al., 2009; Lerman and McDonald, 2009; Kim and Zhai, 2009). Here the desired property varies from the coverage of product aspects (Carenini et al., 2006; Lerman et al., 2009) to the degree of agreement on aspect-specific sentiment (Lerman et al., 2009; Lerman and McDonald, 2009; Kim and Zhai, 2009).

The other paradigm is aspect-based opinion summarization, which is based on identifying aspects and associating opinion sentiment with them. While initially people use information retrieval techniques to recognize aspect terms and opinion expressions (Hu and Liu, 2004; Popescu et al., 2005), recent work seems to favor generative statistical models more (Mei et al., 2007; Lu and Zhai, 2008; Titov and McDonald, 2008b,a; Blei and McAuliffe, 2010; Brody and Elhadad, 2010; Mukherjee and Liu, 2012; Sauper and Barzilay, 2013). One typical problem with these models is that many discovered aspects are not meaningful to end-users. Some of these studies focus on distinguishing aspects in terms of sentiment variation by modeling aspects together with sentiment (Titov and McDonald, 2008a; Lu and Zhai, 2008; Mukherjee and Liu, 2012; Sauper and Barzilay, 2013). However, little attention is given to differentiating review content directly regarding their utilities in review exploration. Mukherjee and Liu (2012) attempted to address this issue by introducing user-provided aspect terms as seeds for learning review aspects, though this approach might not be easily generalized to other domains, as users' point of interest could vary with the review domain. In this paradigm of review summarization, the focus is sentiment-oriented aspect extraction and the output is usually a set of topics words plus their representative text units (Hu and Liu, 2004; Zhuang et al., 2006). Such kind of table-style summaries are often visualized in to charts or graphs to emphasize the summary statistics in an intuitive way which especially suit summarization applications on mobile platforms (Huang et al., 2012). However, such a topic-based summarization framework is beyond the focus of our work, as we aim to adapt traditional extractive techniques to the review domain by introducing review helpfulness ratings as guidance.

In the literature, review helpfulness has been used to facilitate downstream applications such as review recommendation (Dong et al., 2013) and summarization (Liu et al., 2007). However, the helpfulness information has been used only as filtering criteria during input preprocessing. In contrast, our proposed summarization framework further uses review helpfulness as sentence scoring features for content selection. Also, when used for filtering, helpfulness prediction is modeled as classification tasks (binary classification (Dong et al., 2013) or multi-class classification (Liu et al., 2007)), while we consider it as a regression (ranking) problem, which differentiates review helpfulness in finer-grain. Furthermore, Liu et al. (2007) trained the helpfulness classifier based on their expert-provided helpfulness ratings using a specialized rubric that focuses on topic coverage and review's level of detail.

In the presented work, we utilize review helpfulness via using sLDA. The idea of using sLDA in text summarization is not new. However, the model is previously applied at the sentence level (Li and Li, 2012) for query focused multi-document summarization, which requires human annotation on sentence importance with respect to whether a sentence answers the given query. In comparison, our use of sLDA is at the document (review) level, using existing metadata of the document (review helpfulness ratings) as the supervision, and thus requiring no annotation at all.

5.2 HELPFULNESS-GUIDED CONTENT SELECTION

5.2.1 Review-level filtering

The most straight forward way to utilize review helpfulness is filtering out unhelpful reviews based on helpfulness ratings which motivates prior studies of automatic helpfulness prediction (Kim et al., 2006). Early work on Amazon product reviews (Liu et al., 2007) shows that filtering out unhelpful reviews (using a classifier trained on expert-annotated helpfulness ratings) before applying (opinion) summarization yields more positive and negative supporting sentences, and the summary result is more consistent with editor's review by ranking products based on sentiment scores derived from the summary text. In our study, we omit the automated prediction (Xiong and Litman, 2011a) and filter reviews by their helpfulness gold-standard directly. We first calculate the average helpfulness ratings for each domain across all reviews, and consider reviews of helpfulness ratings below the domain-average as unhelpful ones.

5.2.2 Helpfulness-guided sentence scoring

While the most straightforward way to utilize review helpfulness for content selection is through filtering (Liu et al., 2007), we also propose to take into account review helpfulness during sentence scoring by learning helpfulness-related review topics in advance. Because sLDA learns the utility of the topics for predicting review-level helpfulness ratings (decomposing review helpfulness ratings by topics), we develop novel features (*rHelpSum* and *sHelpSum*) based on the inferred topics of the words in a sentence to capture helpfulness in various perspectives. We later use the features for sentence scoring in a helpfulness-guided summarizer (Section 5.3.3).

Compared with LDA (Blei et al., 2003), sLDA (Blei and McAuliffe, 2010) introduces a response variable $y_i \in Y$ to each document D_i during topic discovery. The model not only learns the topic assignment $z_{1:N}$ for words $w_{1:N}$ in D_i, it also learns a function from the posterior distribution of z in D to Y. When Y is the review-level helpfulness gold-standard, the model learns a set of topics predictive of review helpfulness, as well as the utility of z in predicting review helpfulness y_i, denoted as η. (Both z and η are K-dimensional.)

At each inference step, sLDA assigns a topic ID to each word in every review. $z_l = k$ means that the topic ID for word at position l in sentence s is k. Given the topic assignments $z_{1:L}$ to words $w_{1:L}$ in a review sentence s, we estimate the contribution of s to the helpfulness of the review it belongs to (Formula 5.1), as well as the average topic importance in s (Formula 5.2). While $rHelpSum$ is sensitive to the review length, $sHelpSum$ is sensitive to the sentence length.

$$rHelpSum(s) = \frac{1}{N} \sum_{l=1}^{l=L} \sum_{k} \eta_k p(z_l = k) \tag{5.1}$$

$$sHelpSum(s) = \frac{1}{L} \sum_{l=1}^{l=L} \sum_{k} \eta_k p(z_l = k) \tag{5.2}$$

As the topic assignment in each inference iteration might not be the same, Riedl and

Biemann (2012) proposed the *mode* method in their application of LDA for text segmentation – use the most frequently assigned topic for each word in all iterations as the final topic assignment – to address the instability issue. Inspired by their idea, we also use the *mode* method to infer the topic assignment in our task, but only apply the *mode* method to the last 10 iterations.[3] More details about how the model is trained are provided in Section 5.3.2. Note that except for the mode technique, $sHelpSum(s)$ is equivalent to Formula 4.4 in Chapter 4.

5.3 EXPERIMENTAL SETUP

To investigate the utility of exploiting user-provided review helpfulness ratings for content selection in extractive summarization, we develop two helpfulness-guided summarizers based on the MEAD framework (HelpfulFilter and HelpfulSum). In particular, we would like to examine the following hypotheses.

H6 User-provided review helpfulness assessment can be used to improve summarization performance.

H7 Review helpfulness can be automatically predicted at the sentence level.

H8 Using sentence-level review helpfulness information in addition to review-level helpfulness ratings yields better review summarizers.

For Hypothesis **H6**, we compare our systems' performance against a strong unsupervised extractive summarizer that MEAD supports, as our baseline (MEAD+LexRank). To test Hypothesis **H7**, we consider the summarization task as the extrinsic evaluation of our sentence-level review helpfulness predictor, by showing the value of sentence-level helpfulness predictions for summarization content selection, we indirectly validate the model's

[3] As we observed that the topic distribution is usually not well learned at the early stage during the training step when we construct the helpfulness-related review topics in the previous chapter.

prediction performance. With respect to Hypothesis **H8**, we compare HelpfulSum to HelpfulFilter. Note that both summarizers use review-level helpfulness for filtering, while HelpfulSum uses sentence-level helpfulness for sentence scoring without any traditional scoring features used in HelpfulFilter. If HelpfulSum outperforms HelpfulFilter, we can conclude that our Hypothesis **H8** is true.

To focus on sentence scoring only, we use the same MEAD word-based MMR (Maximal Marginal Relevance) reranker (Carbonell and Goldstein, 1998) for all summarizers, and set the length of the output to be 200 words.

Because the target audience for online customer reviews is different from educational peer reviews, we conduct separate evaluation user studies for customer reviews – including camera reviews and movie reviews (Section 5.4), and peer reviews (Section 5.5). For customer reviews, the summarization is performed on each review item (product/movie), and any potential customer can help us judge whether the summary of the product/movie is informative. While for educational peer reviews, the summarization is performed on the reviews of each each paper, and thus only the author of the paper can have a fair judgement on the helpfulness of a review summary.

5.3.1 Data

5.3.1.1 Customer reviews Two domains are examined in the first user study: Camera reviews and Movie reviews. Both corpora were used in the previous chapter of automatically predicting review helpfulness, in which every review has at least three helpfulness votes. Recall that the average helpfulness rating of camera reviews is .80 and that of movie reviews is .74.

Summarization test sets. To create the test sets for summarization evaluation, we randomly sample 18 reviews for each review item (a camera or movie) and randomly select 3 items for each review domain. In total there are 6 summarization test sets (3 items × 2 domains), where each contains 18 reviews to be summarized (i.e. “summarizing 18 camera reviews for Nikon D3200"). In the summarization test sets, the average number of sentences per review is 9 for camera reviews, and 18 for movie reviews; the average number of words per sentence in the camera reviews and movie reviews are 25 and 27, respectively.

5.3.1.2 Peer reviews Physics2014 is collected for our summarization evaluation on peer reviews from a Physics lab in 2014 at the University of Pittsburgh. It contains 6203 peer reviews and the average student-helpfulness rating is .84.[4] Note that the student-helpfulness ratings are skewed towards 1 (after scaling), which is also observed in the other peer review data set (History2008).

Summarization test sets. For educational peer reviews, we create a test set for each student who participated in our second evaluation user study (Section 5.5), by collecting all the reviews the student received.[5] In total there are 37 summarization test sets on peer reviews in this study. In this summarization test set, the average number of sentences per review is 2; the average number of words per sentence is 18.

5.3.2 sLDA training

We implement sLDA based on the topic modeling framework of Mallet (McCallum, 2002) using 20 topics ($K = 20$) and the best hyper-parameters (topic distribution priors α and word distribution priors β) that we learned in our pilot study on LDA. [6]

To learn the topic assignment for each review word, we use all reviews (4050 reviews for camera, 280 reviews for movie, and 6203 for peer) to train the sLDA model for each domain independently. We realize that using a topic model trained without the summarization test sets is more desirable in practice, however, we use all available review-level helpfulness

[4] No expert ratings for this data set.

[5] Every student received 6-30 peer reviews (average = 21) in Physics2014.

[6] In our pilot study, we experimented with various hyper-parameter settings, and trained the model with 100 sampling iterations in both the **E**stimation and the **M**aximization steps. As we found the best results are more likely to be achieved when $\alpha = 0.5$, $\beta = 0.1$, we use this setting to train the sLDA model in our summarization experiment.

information in this experiment to make our best guess on the topic assignment to words in each review sentence to be summarized. Note that our approach is still unsupervised as we do not have gold-standard for sentence scoring or summarization directly.

5.3.3 Three summarizers

Baseline (MEAD+LexRank): The default feature set of MEAD includes *Position*, *Length*, and *Centroid*. Here *Length* is a word-count threshold, which gives score 0 to sentences shorter than the threshold. As we observe that short review sentences sometimes can be very informative as well (e.g., "This camera is so amazing!", "The best film I have ever seen!"), we adjust *Length* to 5 from its default value 9. MEAD also provides scripts to compute *LexRank* (Erkan and Radev, 2004), which is a more advanced feature using a graph-based algorithm for computing relative importance of textual units. We supplement the default feature set with *LexRank* to get the best summarizer from MEAD, yielding the sentence scoring function $F_{baseline}(s)$, in which s is a given sentence and all features are assigned equal weights (same as in the other two summarizers).

$$F_{baseline}(s) = \begin{cases} Position + Centroid + LexRank & \text{if } Length \geq 5 \\ 0 & \text{if } Length < 5 \end{cases} \tag{5.3}$$

HelpfulFilter: This summarizer is a direct extension of the baseline, which considers review-level helpfulness ratings ($hRating$) as an additional filtering criteria in its sentence scoring function $F_{HelpfulFilter}$. (In our study, we omit the automated prediction (Kim et al., 2006; Liu et al., 2008) and filter reviews by their helpfulness gold-standard directly.) We set the cutting threshold to be the average helpfulness rating of all the reviews that we used to train the topic model for the corresponding domain ($hRatingAve(domain)$).

$$F_{HelpfulFilter}(s) = \begin{cases} F_{baseline}(s) & \text{if } hRating(s) \geq hRatingAve(domain) \\ 0 & \text{if } hRating(s) < hRatingAve(domain) \end{cases} \tag{5.4}$$

HelpfulSum: To isolate the contribution of review helpfulness, the second summarizer only uses helpfulness related features in its sentence scoring function $F_{HelpfulSum}$. The features are *rHelpSum* – the contribution of a sentence to the overall helpfulness of its corresponding review, *sHelpSum* – the average topic weight in a sentence for predicting the overall helpfulness of the review (Formulas 5.1 and 5.2), plus *hRating* for filtering. Note that there is no overlap between features used in the baseline and HelpfulSum, as we wonder if the helpfulness information alone is good enough for discovering salient review sentences.

$$F_{HelpfulSum}(s) = \begin{cases} rHelpSum(s) + sHelpSum(s) & \text{if } hRating(s) \geq hRatingAve(domain) \\ 0 & \text{if } hRating(s) < hRatingAve(domain) \end{cases} \tag{5.5}$$

5.4 EVALUATION ON CUSTOMER REVIEWS

For evaluation, we will first present our human evaluation user study. We then will present the automated evaluation result based on a summarization gold-standard collected during the human evaluation study.

5.4.1 Human evaluation

The goal of our human evaluation is to compare the effectiveness of 1) using a traditional content selection method (MEAD+LexRank), 2) using the traditional method enhanced by review-level helpfulness filtering (HelpfulFilter), and 3) using sentence helpfulness features estimated by sLDA plus review-level helpfulness filtering (HelpfulSum) for building an extractive multi-document summarization system for online reviews. Therefore, we use a within-subject design in our user study for each review domain, considering the *summarizer* as the main effect on human evaluation results.

The user study is carried out in the form of online surveys (one survey per domain) hosted by *Qualtrics*[7]. In total, 36 valid users participated in our online-surveys.[8] We randomly assigned 18 of them to the camera reviews, and the rest to the movie reviews.

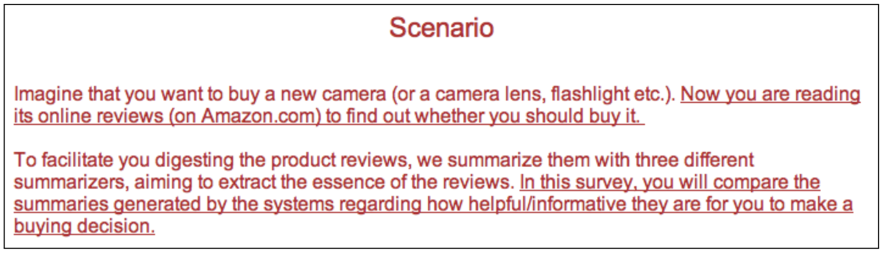
5.4.1.1 Experimental procedures Each online survey contains three summarization sets. The human evaluation on each one is taken in three steps:

Step 1: We first require users to perform **manual summarization**, by selecting 10 sentences from the input reviews (displayed in random order for each visit). This ensures that users are familiar with the input text so that they can have fair judgement on machine-generated results. To help users select the sentences, we provide an introductory scenario at the beginning of the survey to illustrate the potential application in accordance with the domain (e.g., Figure 5.1 and Figure 5.2).

Step 2: We then ask users to perform **pairwise comparison** on summaries generated by the three systems. The three pairs are generated in random order; and the left-or-right display position (in Figure 5.4) of the two summaries in each pair is also randomly selected.

[7] URL: http://www.qualtrics.com

[8] All participants are older than eighteen, recruited via university mailing lists, on-campus flyers as well as social networks online. While we originally considered educational peer reviews as a third domain, about half the participants dropped out in the middle of the survey. Thus we only consider the two e-commerce domains in this experiment (Xiong and Litman, 2014), and proposed a separate followup study tailored to the educational context of the pee-review domain.

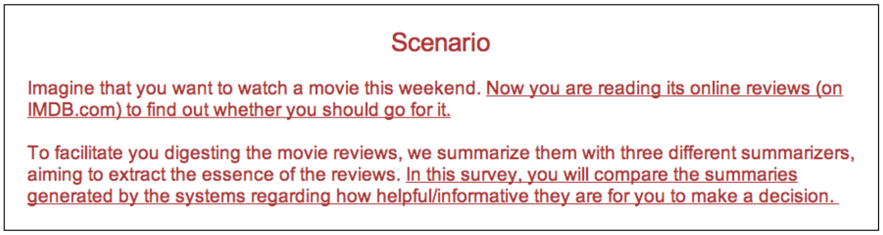


Scenario

Imagine that you want to buy a new camera (or a camera lens, flashlight etc.). Now you are reading its online reviews (on Amazon.com) to find out whether you should buy it.

To facilitate you digesting the product reviews, we summarize them with three different summarizers, aiming to extract the essence of the reviews. In this survey, you will compare the summaries generated by the systems regarding how helpful/informative they are for you to make a buying decision.

Figure 5.1: Scenario for summarizing camera reviews.



Scenario

Imagine that you want to watch a movie this weekend. Now you are reading its online reviews (on IMDB.com) to find out whether you should go for it.

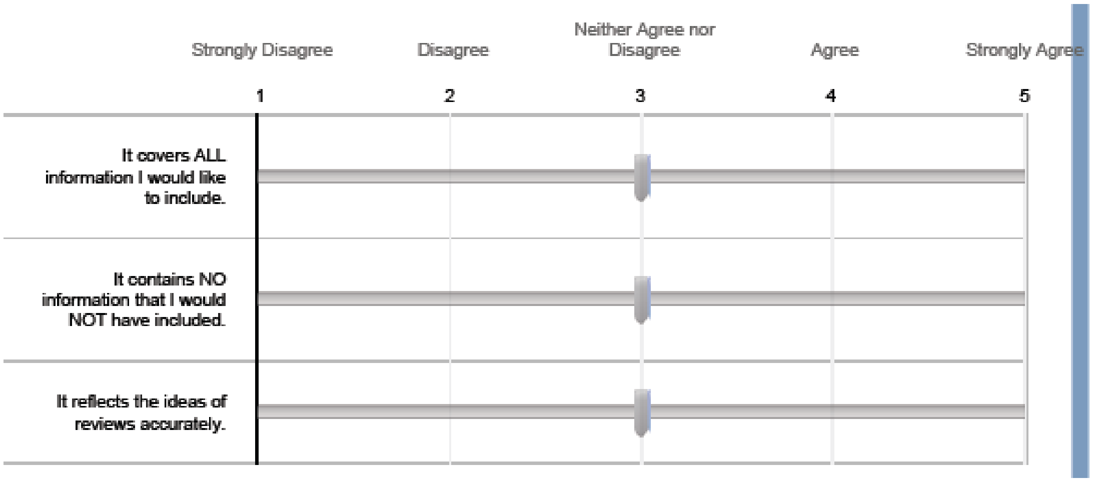
To facilitate you digesting the movie reviews, we summarize them with three different summarizers, aiming to extract the essence of the reviews. In this survey, you will compare the summaries generated by the systems regarding how helpful/informative they are for you to make a decision.

Figure 5.2: Scenario for summarizing movie reviews.

Here we use the same 5-level preference ratings used for pairwise comparison in (Lerman et al., 2009), and translate them into integers from -2 to 2 in our result analysis.

Step 3: Finally, we ask users to evaluate the three summaries in isolation regarding the summary quality in three content-related aspects: *recall*, *precision* and *accuracy* (top, middle and bottom in Figure 5.3, respectively), which were used in (Carenini et al., 2006). In this **content evaluation**, the three summaries are randomly visited and the users rate the proposed statements (one for each aspect) on a 5-point scale.

Complete examples of the survey materials are provided in Appendix , including the summarization test set and the summaries generated by the three summarizers, one for each domain.



It covers ALL information I would like to include.

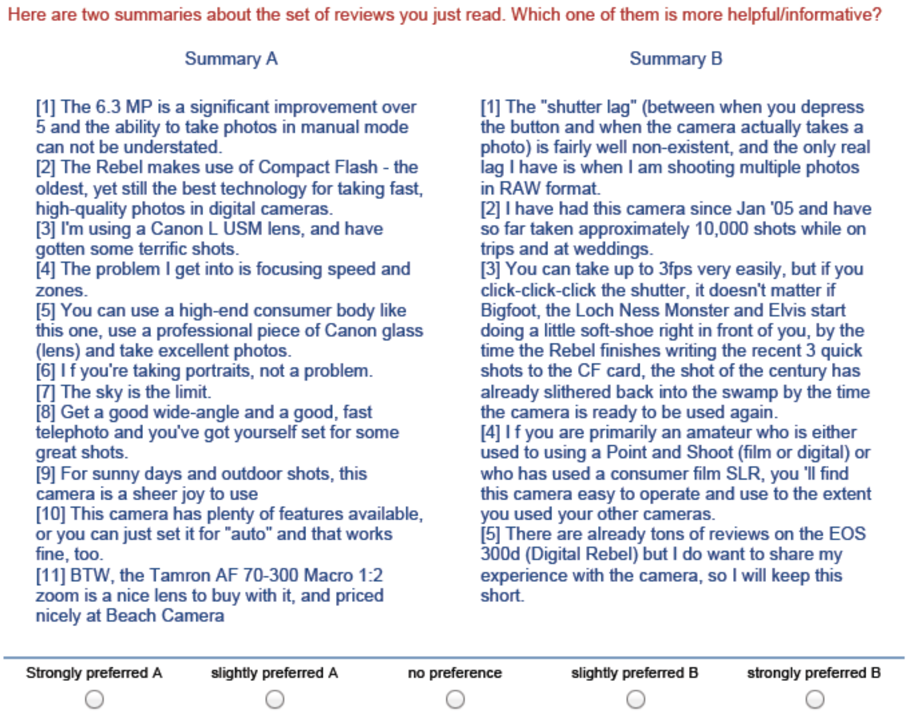
It contains NO information that I would NOT have included.

It reflects the ideas of reviews accurately.

Figure 5.3: Content evaluation on a summary's *recall* (top row), *precision* (middle row) and *accuracy* (bottom row).

5.4.1.2 Results

Pairwise comparison. We use a mixed linear model to analyze user preference over the three summary pairs separately, in which "summarizer" is a within-subject factor, "review item" is the repeated factor, and "user" is a random effect. Results are summarized in Table 5.1. Positive preference ratings on "A over B" means A is preferred over B; negative ratings means B is preferred over A. As we can see, **HelpfulSum** is the best: it is consistently preferred over the other two summarizers across domains and the preference is significant throughout conditions except when compared with HelpfulFilter on movie reviews. **HelpfulFilter** is significantly preferred over the baseline (MEAD+LexRank) for movie reviews, while, surprisingly, the baseline works better than HelpfulFilter for camera reviews. A further look at the compression rate (cRate) of the three systems (Table 5.2) shows that HelpfulFilter generates much shorter sum-



Here are two summaries about the set of reviews you just read. Which one of them is more helpful/informative?

Summary A

[1] The 6.3 MP is a significant improvement over 5 and the ability to take photos in manual mode can not be understated.
[2] The Rebel makes use of Compact Flash - the oldest, yet still the best technology for taking fast, high-quality photos in digital cameras.
[3] I'm using a Canon L USM lens, and have gotten some terrific shots.
[4] The problem I get into is focusing speed and zones.
[5] You can use a high-end consumer body like this one, use a professional piece of Canon glass (lens) and take excellent photos.
[6] I f you're taking portraits, not a problem.
[7] The sky is the limit.
[8] Get a good wide-angle and a good, fast telephoto and you've got yourself set for some great shots.
[9] For sunny days and outdoor shots, this camera is a sheer joy to use
[10] This camera has plenty of features available, or you can just set it for "auto" and that works fine, too.
[11] BTW, the Tamron AF 70-300 Macro 1:2 zoom is a nice lens to buy with it, and priced nicely at Beach Camera

Summary B

[1] The "shutter lag" (between when you depress the button and when the camera actually takes a photo) is fairly well non-existent, and the only real lag I have is when I am shooting multiple photos in RAW format.
[2] I have had this camera since Jan '05 and have so far taken approximately 10,000 shots while on trips and at weddings.
[3] You can take up to 3fps very easily, but if you click-click-click the shutter, it doesn't matter if Bigfoot, the Loch Ness Monster and Elvis start doing a little soft-shoe right in front of you, by the time the Rebel finishes writing the recent 3 quick shots to the CF card, the shot of the century has already slithered back into the swamp by the time the camera is ready to be used again.
[4] I f you are primarily an amateur who is either used to using a Point and Shoot (film or digital) or who has used a consumer film SLR, you 'll find this camera easy to operate and use to the extent you used your other cameras.
[5] There are already tons of reviews on the EOS 300d (Digital Rebel) but I do want to share my experience with the camera, so I will keep this short.

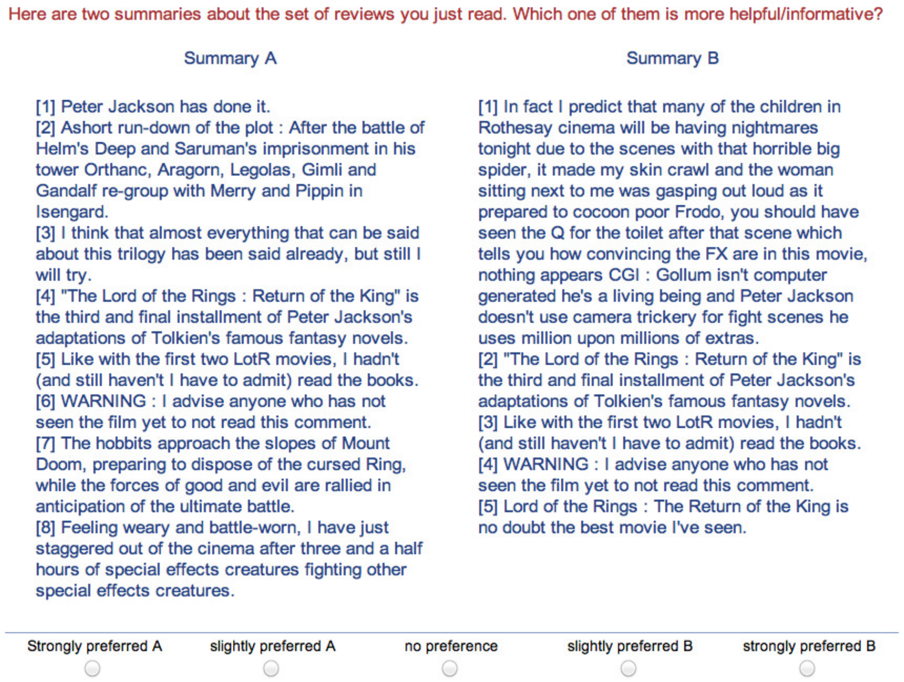
Strongly preferred A	slightly preferred A	no preference	slightly preferred B	strongly preferred B
○	○	○	○	○

Figure 5.4: Example of pairwise comparison for summarizing camera reviews (left:HelpfulSum, right: the baseline).

maries than the other two on camera reviews.[9] As suggested in (Napoles et al., 2011), if system A has better evaluation results than system B but output longer summaries, it is not necessarily the case that system A is better than system B. So the worse performance of HelpfulFilter on camera reviews may due to the average shorter length of the summaries that HelpfulFilter generated for the particular three set of camera reviews.

Content evaluation. We summarize the average quality ratings (Figure 5.3) received

[9] While we limit the summarization output to be 200 words in MEAD, as the content selection is at the sentence level, the summaries can have different number of words in practice. Considering that word-based MMR controls the redundancy in the selected summary sentences ($\lambda = 0.5$ as suggested), there might be enough content to select using $F_{HelpfulFilter}$.



Here are two summaries about the set of reviews you just read. Which one of them is more helpful/informative?

Summary A

[1] Peter Jackson has done it.
[2] Ashort run-down of the plot : After the battle of Helm's Deep and Saruman's imprisonment in his tower Orthanc, Aragorn, Legolas, Gimli and Gandalf re-group with Merry and Pippin in Isengard.
[3] I think that almost everything that can be said about this trilogy has been said already, but still I will try.
[4] "The Lord of the Rings : Return of the King" is the third and final installment of Peter Jackson's adaptations of Tolkien's famous fantasy novels.
[5] Like with the first two LotR movies, I hadn't (and still haven't I have to admit) read the books.
[6] WARNING : I advise anyone who has not seen the film yet to not read this comment.
[7] The hobbits approach the slopes of Mount Doom, preparing to dispose of the cursed Ring, while the forces of good and evil are rallied in anticipation of the ultimate battle.
[8] Feeling weary and battle-worn, I have just staggered out of the cinema after three and a half hours of special effects creatures fighting other special effects creatures.

Summary B

[1] In fact I predict that many of the children in Rothesay cinema will be having nightmares tonight due to the scenes with that horrible big spider, it made my skin crawl and the woman sitting next to me was gasping out loud as it prepared to cocoon poor Frodo, you should have seen the Q for the toilet after that scene which tells you how convincing the FX are in this movie, nothing appears CGI : Gollum isn't computer generated he's a living being and Peter Jackson doesn't use camera trickery for fight scenes he uses million upon millions of extras.
[2] "The Lord of the Rings : Return of the King" is the third and final installment of Peter Jackson's adaptations of Tolkien's famous fantasy novels.
[3] Like with the first two LotR movies, I hadn't (and still haven't I have to admit) read the books.
[4] WARNING : I advise anyone who has not seen the film yet to not read this comment.
[5] Lord of the Rings : The Return of the King is no doubt the best movie I've seen.

Strongly preferred A	slightly preferred A	no preference	slightly preferred B	strongly preferred B
○	○	○	○	○

Figure 5.5: Example of pairwise comparison for summarizing movie reviews (left:HelpfulSum, right: HelpfulFilter).

by each summarizer across review items and users for each review domain in Table 5.3. First, we examine the main effect of the summarizer (3 levels) as a within-subject factor on the content evaluation results. The impact of summarizer is significant only on *recall* and *accuracy* for camera reviews. Post-hoc test shows that HelpfulSum is significantly better than HelpfulFilter (p = .001 for *accuracy*, p = .098 for *recall*) and the baseline is significantly better than HelpfulFilter (p = .003 for *accuracy*, p = .001 for *recall*), but there is no difference between HelpfulSum and the baseline. While for movie reviews, no significant difference is found among the three summarizers on any quality metric, HelpfulSum has the best performance on all metrics regarding the absolute scores.

Pair	**Domain**	**Est. Mean**	**Std. Dev.**	**Sig.**
HelpfulFilter over MEAD+LexRank	Camera	-.602	1.25	.001
	Movie	.621	1.10	.000
HelpfulSum over MEAD+LexRank	Camera	.424	1.22	.011
	Movie	.601	1.05	.000
HelpfulSum over HelpfulFilter	Camera	1.18	1.34	.000
	Movie	.160	1.16	.310

Table 5.1: Mixed-model analysis of user preference ratings (18 subjects × 3 items, $N = 54$) in pairwise comparison across domains. Confidence interval = 95%. The preference rating is ranged from -2 to 2.

Summarizer	**Camera**	**Movie**
MEAD+LexRank	6.07%	2.64%
HelpfulFilter	3.25%	2.39%
HelpfulSum	5.94%	2.69%
Human (Ave.)	6.11%	2.94%

Table 5.2: Compression rate of the three systems across domains.

With respect to pairwise evaluation, content evaluation yields consistent results on camera reviews between HelpfulFilter vs. the baseline and HelpfulSum vs. HelpfulFilter. However, only pairwise comparison (preference ratings) shows significant difference between HelpfulSum vs. the baseline and the difference in the summarizers' performance on movie reviews. Prior work on review summarization evaluation also suggests that pairwise comparison is more suitable than content evaluation for human evaluation (Lerman et al.,

Summarizer	Size	Camera		
		Precision	Recall	Accuracy
MEAD+LexRank	54	2.63(1.10)	**3.24(1.04)**	3.57(.980)
HelpfulFilter	54	**2.78(1.19)**	2.74 (1.20)	3.31(1.11)
HelpfulSum	54	2.41(1.07)	3.19(1.07)	**3.69(.948)**

Summarizer	Size	Movie		
		Precision	Recall	Accuracy
MEAD+LexRank	54	2.50(1.07)	2.59(1.11)	2.93(1.04)
HelpfulFilter	54	2.44(.101)	2.61(1.23)	2.96(1.11)
HelpfulSum	54	**2.52(.104)**	**2.67(1.15)**	**3.02(1.10)**

Table 5.3: Average human ratings for content evaluation (Standard Deviation within parentheses). The best result on each metric is bolded for every review domain (the higher the better).

2009).

5.4.2 Automated evaluation based on ROUGE metrics

Although human evaluation is generally preferred over automated metrics for summarization evaluation, we report our automated evaluation results based on ROUGE scores (Lin, 2004) using references collected from the user study. ROUGE stands for recall-oriented understudy for gisting evaluation, containing a set of metrics by examining the overlapped text units such as n-gram and word sequence between a test summary and the corresponding reference(s).

For each summarization test set, we have 3 machine generated summaries and 18 human

Summarizer	**R-1**	**R-2**	**R-SU4**
MEAD+LexRank	.333	.117	.110
HelpfulFilter	.346	**.121**	**.111**
HelpfulSum	**.350**	.110	.101
Human	.360	.138	.126

Table 5.4: ROUGE evaluation on camera reviews.

Summarizer	**R-1**	**R-2**	**R-SU4**
MEAD+LexRank	.281	.044	.047
HelpfulFilter	.273	.040	.041
HelpfulSum	**.325**	**.095**	**.090**
Human	.339	.093	.093

Table 5.5: ROUGE evaluation on movie reviews.

summaries. We compute the ROUGE scores in a leave-1-out fashion: for each machine generated summary, we compare it against 17 out of the 18 human summaries and report the score average across the 17 runs; for each human summary, we compute the score using the other 17 as references, and report the average human summarization performance.

Evaluation results are summarized in Table 5.4 and Table 5.5, in which we report the F-measure for R-1 (unigram), R-2 (bigram) and R-SU4 (skip-bigram with maximum gap length of 4)[10], following the convention in the summarization community. Here we observe slightly different results with respect to human evaluation: for camera reviews,

[10] Because ROUGE requires all summaries to have equal length (word counts), we only consider the first 100 words in every summary.

the difference among the three machine-summarizers is not significant, while HelpfulSum achieves the best R-1 score and HelpfulFilter works best regarding R-2 and R-SU4 based on the absolute values. In both cases the baseline is never the best. For movie reviews, summarizer has significant impact ($p < .05$): HelpfulSum significantly outperforms the other machine-summarizers on all ROUGE measurements, and the improvement is over 100% on R-2 and R-SU4, similar to average human performance. This is consistent with the result of pairwise comparison in that HelpfulSum works better than both HelpfulFilter and the baseline on movie reviews.

5.4.3 Human summary analysis

To get a comprehensive understanding of the challenges in extractive review summarization, we analyze the agreement in human summaries collected in our user study at different levels of granularity, regarding heuristics that are widely used in existing extractive summarizers.

Average word/sentence counts. Figure 5.6 illustrates the trend of average number of words and sentences shared by different number of users across review items for each domain. As it shows, no sentence is agreed by over 10 users, which suggests that it is hard to make humans agree on the informativeness of review sentences. Prior analysis on news articles (Lin and Hovy, 2002) also report low inter-human agreement in sentence selection, though the reported coverage is 29% between two judges, which is comparatively better than what we observed for online reviews.

Word frequency We then compute the average probability of words (in the input) used by different number of human summarizers to see if the word frequency pattern found in news articles (words that human summarizers agreed to use in their summaries are of high frequency in the input text (Nenkova and Vanderwende, 2005)) holds for online reviews. Figure 5.7 confirms this. However, the average word probability is below 0.01 in those shared by 14 out of 18 summaries[11]; the flatness of the curve seems to suggest that word

[11] The average probability of words used by all 4 human summarizers are 0.01 across the 30 DUC'03

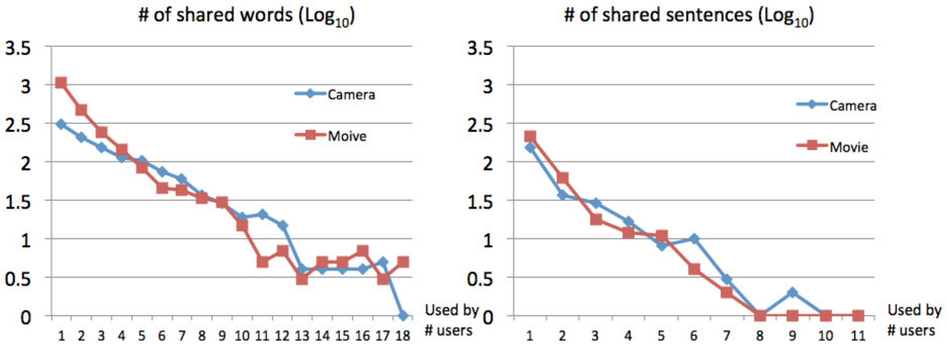


Figure 5.6: Average number of words (w) and sentences (s) in agreed human summaries.

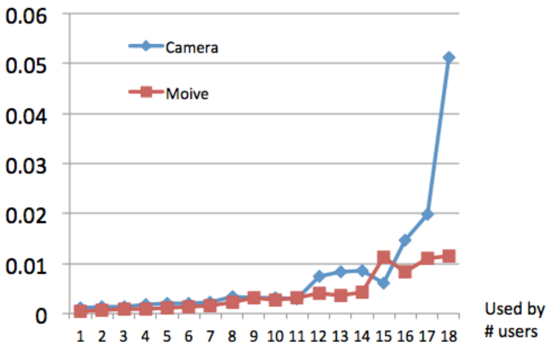


Figure 5.7: Average probability of words used in human summaries.

frequency alone is not enough for capturing the salient information in input reviews.

KL-divergence. Another widely used heuristic in multi-document summarization is minimizing the distance of unigram distribution between the summary and the input text (Lin et al., 2006). We wonder if this applies to online review summarization. For each testing

sets(Nenkova and Vanderwende, 2005).

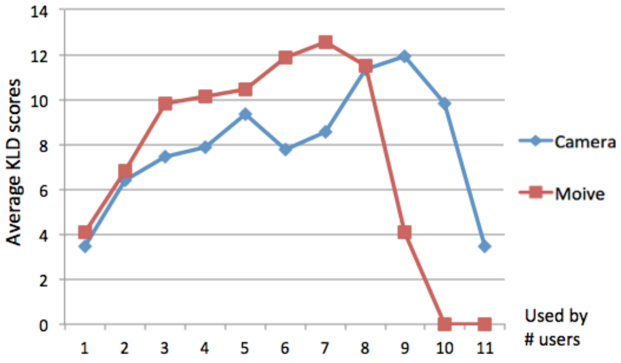


Figure 5.8: Average KL-Divergence between input and sentences used in human summaries.

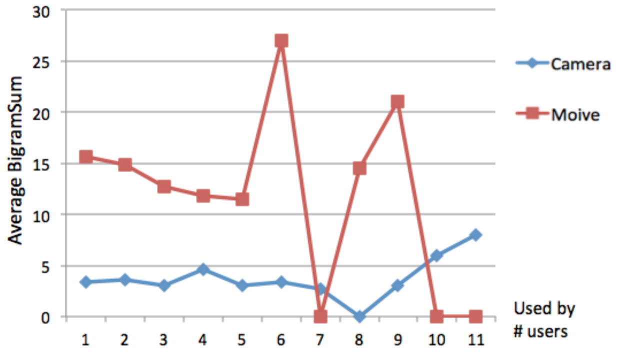


Figure 5.9: Average BigramSum of sentences used in human summaries.

set, we group review sentences by the number of users who selected them in their summaries, and compute the KL-divergence (KLD) between each sentence group and the input. The average KL-divergence of each group across review items are visualized in Figure 5.8, showing that this intuition is incorrect for our review domains. Actually, the pattern is quite the opposite, especially when the number of users who share the sentences is less than 8. Thus traditional methods that aim to minimize KL-divergence might not work well for online reviews.

Bigram coverage. Recent studies proposed a simple but effective criteria for extractive summarization based on bigram coverage (Nenkova and Vanderwende, 2005; Gillick and Favre, 2009). The coverage of a given bigram in a summary is defined as the number of input documents the bigram appears in, and presumably good summaries should have larger sum of bigram coverage (BigramSum). However, as shown in Figure 5.9, this criteria might not work well in our case either. For instance, the BigramSum of the sentences that are shared by 3 human judges is smaller than those shared by 1 or 2 judges.

To conclude, our evaluation results on customer reviews support our hypotheses about the value of both review-level and sentence-level review helpfulness for building effective summarization systems for both camera reviews and movie reviews. The effectiveness of HelpfulSum also shows that our helpfulness-related topics learned by sLDA can different review helpfulness at the sentence level.

5.5 EVALUATION ON EDUCATIONAL PEER REVIEWS

When summarizing educational peer reviews, the application scenario is different: the target users are students who received the reviews, namely, the paper authors. As each peer review is sent to only one student, there is only one target user for each review summary. In this case, we need to apply summarization for each student separately, querying their feedback on the summarization of the peer reviews that they received.

Therefore, to evaluate the proposed summarization idea on educational peer reviews, we conduct a separate user study with students of the Physics lab from which we collected Physics2014. In particular, we treat each student's received peer reviews as a summarization test set, and setup an online survey for each student. We advertised the user study to the students by email, with five dollar amazon gift cards as a reward.[12]

If a student agreed to participate in our experiment, we directed them to their relevant survey hosted on Qualtrics.com. In total 37 out of 304 students from the Physics lab participated in our study, yielding 37 summarization test sets for human evaluation.

5.5.1 Experimental procedure

In this experiment, we remove the manual summarization step (which was originally used to familiarize participants with the summarization input), as the students are supposed to read the reviews already. Thus there is no reference for us to evaluate the summarization performance automatically based on ROUGE scores. Therefore, in this section, we only present our human evaluation on peer reviews. The peer-review human evaluation is conducted in a similar way as we did for customer reviews in Section 5.4.1.

In this user study, every survey is for a particular student (evaluating a different summarization test set) which has two parts: the first part contains 3 pairwise comparison questions between the three summarizers; the second part contains 3 content evaluation questions, one each summarizer. Note that in the previous experiment, participants assigned to the same domain evaluated the same three sets of three summarization results, however, in the peer review experiment, every student examined only one summarization set, which is different from one student to another.

5.5.2 Main effect of the summarizer

First we analyze student responses with the summarizer as the only main effect, as we did in the previous user study. However, we only observe that the baseline has significantly higher *precision* than helpfulFilter.

[12] We paid ten dollars to the participant in the previous user study, as the previous one involves manual summarization which take much more time. The previous user study generally took 30 60 minutes, while the majority of students finished this study in five minutes.

Pairwise comparison. We present the estimated mean of the preference ratings of each summarizer pair across 37 participants in Table 5.6. For pair "A over B", positive ratings mean that A is preferred over B. Because different students examined different review sets, no repeated factor is involved in the result analysis. Therefore, we use one sample T-test against 0 (no preference) instead of the mixed model for statistical analysis. However, no significant preference is observed between any summarizer pair (all p values are greater than .05). Based on the absolute values, HelpfulFilter is preferred over the baseline and HelpfulSum is preferred over HelpfulFilter. However, HelpfulSum is rated worse when directly compared against the baseline. Compared with our preference comparison result on camera reviews and movie reviews (Table 5.1), the preference between the three summarizers is subtle on educational peer reviews (Table 5.6).

Pair	**Est. Mean**	**Std. Dev.**	**Sig.**
HelpfulFilter over MEAD+LexRank	.054	1.03	.750
HelpfulSum over MEAD+LexRank	-.135	1.03	.431
HelpfulSum over HelpfulFilter	.027	.80	.838

Table 5.6: Result analysis of user preference ratings ($N = 37$) in pairwise comparison on educational peer reviews. One sample T-test is performed against 0. Confidence interval = 95%. The preference rating is ranged from -2 to 2.

Content evaluation. Average ratings of each of the three summarizers on the content evaluation metrics (Precision, Recall and Accuracy) are summarized in Table 5.7. While the baseline received the highest rating on precision, HelpfulSum is best on *recall* and HelpfulFilter is best on the summary content *accuracy*. But statistical test finds no significant difference among the three summarizers except for *precision*, on which we observed that the impact of summarizer is in trend ($p = .07$).

Summarizer	Size	Peer review		
		Precision	Recall	Accuracy
MEAD+LexRank	37	**3.14(1.00)**	3.35(.920)	3.68(.818)
HelpfulFilter	37	2.73(.902)	3.11(.966)	**3.74(.534)**
HelpfulSum	37	2.84(.986)	**3.41(1.01)**	3.62(.828)

Table 5.7: Average human ratings for content evaluation (Standard Deviation in parentheses). Best results on each metric are bolded (the higher the better).

5.5.3 Impact of participant's demographic factors

As our prior work on evaluating topic-word review analysis in a visual analytic tool developed for evaluating student peer review performance (Xiong and Litman, 2013) showed that user background factors do influence the utility of topic-word analytics (participants who have prior teaching experience or peer-review experience tend to have better performance on our user study tasks and higher satisfaction towards the proposed analytic approach), we suspect that student demographic features affect how they perceive the utility of different summarizers on peer reviews as well. Therefore, we further investigate the impact of student demographic information and student performance on the summarization human evaluation result. These factors include:

- Teaching experience ($expTA$) – 1: has teaching experience, 0: No.
- *Gender* – 1: male, 2: female.
- *NativeSpeaker* – 1: English 0: others.

We also look at student peer-review performance, both as a paper author and a reviewer, which includes:

- *Rating* – The average paper rating that a student receives. (Each paper is rated on multiple dimensions as specified by the instructor on a scale from 1 to 7.)
- *Helpfulness* – The average of student-provided helpfulness ratings that a student received as a reviewer. (1 - 7)
- *ReviewingAccuracy* – The average reviewing accuracy rating that a student received. The accuracy ratings are automatically generated by SWoRD. (0 - 1)
- *Time* – the time spent on the user study (*Time*), measured by minutes.

Among the 37 participants, 17 are male and 20 are female; 12 have teaching experience; only 2 are English non-native speakers. The average paper rating is 5.54; the average helpfulness and reviewingAccuracy is 4.37 and .73 respectively.

These factors are examined on the difference between the three summarizers, pairwisely, with respect to the human evaluation measurements, which include the three preference ratings (denoted as preference(x, y)) and 9 rating differences (denoted as eval(x, y)) between each two of the three summarizers on the 3 content evaluation metrics:

- $Preference(X, Y)$. Preference of X over Y. The ratings collected in pairwise comparison.
- $Eval(X, Y) = Eval(X) - Eval(Y)$. *Eval* can be *precision*, *recall* or *accuracy*. $Eval(X)$ is the evaluation result collected in content evaluation.

5.5.3.1 Automatic linear modeling analysis

Firstly, we run stepwise linear regression on each of the dependent variables with the 7 factors using the linear modeling procedure provided by SPSS.

The automatic linear modeling procedure in SPSS automatically transforms features to the same scale, identifies the outliers with respect to Cook's Distance, and excludes the outliers in model building. It uses forward stepwise feature selection and compares model performance based on information criterion (models with smaller values fit better). The analysis result (in Table 5.8) shows that student language background (*nativeSpeaker*),

Measurement	**# of outliers**	**Sig. factor ($p < .05$)**	**Coef.**	**Sig.**
Preference(HelpfulFilter, baseline)	2			
Preference(HelpfulSum, baseline)	2	nativeSpeaker	1.7	.019
Preference(HelpfulSum, HelpfulFilter)	0			
Precision(HelpfulFilter, baseline)	1	Rating	- .6	.043
Precision(HelpfulSum, baseline)	1	Rating	-.75	.039
Precision(HelpfulSum, HelpfulFilter)	6			
Recall(HelpfulFilter, baseline)	4	expTA	.73	.033
Recall(HelpfulSum, baseline)	1			
Recall(HelpfulSum, HelpfulFilter)	3	expTA	-.91	.015
Accuracy(HelpfulFilter, baseline)	3			
Accuracy(HelpfulSum, baseline)	1	Time	-91	.021
Accuracy(HelpfulSum, HelpfulFilter)	3	expTA Time	-.45 -.106	.030 .008

Table 5.8: Automatic linear modeling analysis of the significant student-related factors on summarization human evaluation ratings (N=37).

teaching experience ($expTA$), writing performance (*Rating*) and the time spent on the survey ($Time$) are significant factors in explaining the human evaluation difference between the three summarizers. A further look at the trend of how each of the corresponding dependent variables correlates with the significant factor suggests: 1) non-native speakers prefer HelpfulSum to the baseline, 2) Students who received lower ratings on their papers rate both HelpfulFilter and helpfulSum higher than the baseline on precision, and 3) students who have no teaching experience are more likely to rate HelpfulFilter higher than HelpfulSum on content recall and accuracy, and these students think that HelpfulFilter generates summaries with better recall than the baseline.

Here we see that low-performance students think that the summaries generated by the helpfulness-guided summarizers have more information that they would like to see (higher precision) compared with the baseline. Students with no teaching experience ($expTA = 0$) favor the filtering approach with helpfulness rating given by themselves for recall concerns ($expTA = 0$ has positive coefficient on $\mathit{Recall}(HelpfulFilter, \mathit{baseline})$ and negative coefficient on $\mathit{Recall}(HelpfulSum, HelpfulFilter)$). This suggests that user-provided helpfulness ratings capture useful feedback that is not recognized by the traditional summarization method, at least from the perspective of students who have no teaching experience. However, for students who have teaching experience ($expTA = 1$) HelpfulSum seems to capture the point of interest more correctly than HelpfulFilter. In terms of $Time$, it seems that the longer the students spent on the evaluation task the more likely they would identify problems with HelpfulSum.

5.5.3.2 Mixed model analysis of student group-differences

To better understand the group difference regarding student profile in judging the effectiveness of the proposed summarizers, we analyze the fixed effects of student performance (excluding $Time$) and student demographic information using mixed models. For the whole class (304 students), $mean(\mathit{Rating}) = 5.50$, $mean(Helpfulness) = 4.37$, $mean(\mathit{reviewAccuracy}) = .69$. For the subset of the 37 participants, $mean(\mathit{Rating}) = 5.54$, $mean(Helpfulness) = 4.37$, $mean(\mathit{reviewAccuracy}) = .73$. Because this analysis only handles categorical factors, we convert the numeric variables to binary categorical variables by comparing them to the variable mean of the whole class (304 students). If x is greater than the mean, then x is converted to 1 labeled as “high”, else x is converted to 0 labeled as “low”. The mean value of the transformed variables are provided in Table 5.9. Table 5.9 also shows that the majority of our participants are high performance students.

Factor	Original mean	Original Std. D	Transformed mean	# of low	# of high
Rating	5.54	.58	.59	15	22
Helpfulness	4.37	.41	.65	13	24
ReviewingAccuracy	.73	.15	.62	14	23

Table 5.9: Data transformation for mixed model analysis on student-related factors.

For analysis, first we run a general linear model with the six factors as between-subject factors and summarizer-pair (3) and measurement (4) as within-subject factors. No significant within-subject effect is observed. Therefore, we consider the 12 dependent variables as independent with each other and test the six factors on them separately using ANOVA test. Significant group differences are summarized in Table 5.10.

As show in Table 5.10, students who are non-native speakers ($nativeSpeaker = 0$) prefer HelpfulSum over the baseline ($F(1, 35) = 5.06$ $p = .019$). From the perspective of students who **have** teaching experience ($expTA = 1$), HelpfulSum is better than the baseline ($F(1, 35) = 8.78$, $p = .005$) and the baseline is better than HelpfulFilter ($F(1, 35) = 7.07$, $p = .012$) on recall. And these students also prefer HelpfulSum over HelpfulFilter in terms of summary content accuracy ($F(1, 35) = 5.13$, $p = .30$).

With respect to student writing performance, students who receive **low** ratings on their paper ($Rating = low$) rate the proposed summarizers higher than the baseline on the precision of the summarized content ($F(1, 35) = 8.11$, $p = .007$ for "HelpfulFilter over Baseline" and $F(1, 35) = 7.36$ $p = .010$ for "HelpfulSum over Baseline"), which are consistent with the results of the automatic linear modeling analysis. However, none of these factors are found significant on content recall or accuracy (and thus omitted from the table). In addition, the analysis on the binary version of *Rating* shows such difference is also significant in the same way for *recall* ($F(1, 35) = 5.45$, $p = .025$ for "HelpfulFilter over Baseline" and $F(1, 35) = 4.19$ $p = .048$ for "HelpfulSum over Baseline").

When multiple factors are found significant on one measurement, we further consider them together including their interactions using a mixed model analysis. However, no significant interaction is ever observed.

Factor	**Level**	**Size**	**Measurement**	**M**	**SD**
nativeSpeaker	1	25	Preference(HelpfulSum, baseline)	-.23	.97
	0	2		1.5	.71
expTA	1	12	Accuracy(HelpfulSum, HelpfulFilter)	.17	.39
	0	25		-.32	.69
	1	12	Recall(HelpfulFilter, baseline)	-.83	.94
	0	25		.04	.94
	1	12	Recall(HelpfulSum, HelpfulFilter)	1.0	1.0
	0	25		.04	.98
Rating	high	22	Precision(HelpfulFilter, baseline)	-.77	.11
	low	15		.13	.60
	high	22	Precision(HelpfulSum, baseline)	-.73	.12
	low	15		.33	1.2
	high	22	Recall(HelpfulFilter, baseline)	-.55	.97
	low	15		.20	.94
	high	22	Recall(HelpfulSum, baseline)	.05	.38
	low	15		.07	.80

Table 5.10: Analysis of student group differences regarding their demographic background on summarization human evaluation ratings.

The analysis presented above reveals the limitation and risk of using student provided helpfulness ratings as reference for picking helpful peer reviews for certain students. While frequently mentioned feedback is perceived important (captured by the baseline), the user study still shows that user (student) provided review helpfulness can be useful, especially in the opinion of low-performance students ($rating = low$) and students who have teaching experience ($expTA = 1$). Since there are only two non-native English speakers, we need more data to verify whether language background is a confounding factor in how students perceive the utility of the helpfulness-guided summarizers in general.

With respect to student teaching experience, while students who have not taught before consider HelpfulFilter is the best among the three summarizers for recall, students who have teaching experience consider HelpfulSum is better than HelpfulFilter for both recall (for $expTA = 1$, the mean of $Recall(HelpfulSum, HelpfulFilter)$ is 1.0, the mean of $Accuracy(HelpfulSum, HelpfulFilter)$ is .17) and accuracy. This indicates that the helpfulness information provided by students who have teaching experience can generate helpful sentence-level helpfulness features that enables HelpfulSum to outperform HelpfulFilter. This shows the importance of picking good helpfulness gold standard for estimating review helpfulness at the sentence level, which is consistent with our observation in our prior work that helpfulness models trained on student-helpfulness ratings are less predictive compared to the model trained on expert-helpfulness ratings (Xiong and Litman, 2011b).

In summary, our peer-review evaluation user study provides mixed evidence for our hypotheses (**H7** and **H8**). Whether student-provided review helpfulness benefits peer-review summarization depends on the particular student type: adding review-level helpfulness for filtering out unhelpful reviews is preferred by students who have low writing performance; students who have prior teaching experience consider that using sentence-level helpfulness predictions for content selection is useful as it covers more valuable (higher recall) and accurate (higher accuracy) ideas compared to using review-level helpfulness ratings with traditional content selection features.

5.6 DISCUSSION

While the proposed helpfulness-guided summarization framework works for customer reviews as presented in the previous experiment, the value of review helpfulness for summarization, especially in our case of using student-provided helpfulness ratings as the gold-standard, varies with students. In particular, students who have high performance in the corresponding writing assignment prefer the baseline which values frequently mentioned content, while students who have lower performance think it helpful to add the filtering step using the helpfulness rating generated by themselves. Though having no evidence, we suspect that papers written by the high-performance students are generally less problematic; the received reviews are less likely to suggest many revisions. In contrast, low-performance students' papers might be more problematic, leaving larger room for reviewers to comment on various things. In the second case, the student-generated helpfulness meta-data help students to focus on more important ideas (in their own point of view), which makes HelpfulFilter preferred over the traditional summarization method.

In general, we observe smaller preferences between all three summarizers when evaluating on peer reviews compared with applying them on customer reviews. While the helpfulness-guided approach is preferred in most cases on customer review, students' preference for helpfulness depends on student writing performance as well as their teaching experience. Especially, low-performance students and non-expert students like the filtering approach but think the traditional summarization approach is more effective than using sentence-level helpfulness for content selection, high-performance and expert students think using both review-level and sentence-level helpfulness better than using review-level helpfulness alone in terms of content recall and accuracy.

We would like to point out that we set the summary length to be 200 in all our summarization experiments. When configured with a different summary length, the experiment results would be slightly different. When the output summaries are shorter (less number of sentences), the difference between HelpfulSum and the other two would be greater (less likely to catch a similar sentence in two different orders), while the difference between HelpfulFilter and the baseline would be smaller (less likely to encounter a sentence that has a helpful rating lower than the threshold). When the output summaries are long enough, the set of sentences selected by HelpfulSum and HelpfulFilter could be the same, different only in sentence order. In such a case, they are supposed to receive the same content quality ratings.

While we have demonstrated how review helpfulness can be predicted at the sentence level using sLDA, we have not evaluated the performance directly. With the summarization references collected from multiple human subjects during the user study, sentence-level evaluation metrics could be developed based on how well a sentence is covered in all human summaries. These human summaries may also be used to further tune the summarization model, for example, to optimize the feature weights in the sentence-scoring function.

Regarding the computational cost, we realize that training the supervised topic model is time consuming, which prevents HelpfulSum from being used in a real-time operational setting. This would not be a problem if off-line batch-processing is allowed. Also, if the supervised topic model is already well trained, we can use it directly for topic inference in the summarization task, which only takes a few seconds. In this case, we need to make sure that the topic model is trained on reviews similar to the reviews to be summarized.

5.7 SUMMARY

We propose a novel unsupervised extractive approach for summarizing online reviews by exploiting review helpfulness ratings for content selection. We demonstrate that the helpfulness metadata can not only be directly used for review-level filtering, but can also be used as the supervision of sLDA to predict review helpfulness at the sentence level. This approach leverages the existing metadata of online reviews, requiring no annotation and generalizable to multiple review domains. Our experiment based on the MEAD framework shows that HelpfulFilter is preferred over the baseline (MEAD+LexRank) on movie reviews in human evaluation. HelpfulSum, which utilizes review helpfulness at both the review and sentence level, significantly outperforms the baseline in human and automated evaluation for both domains. In the educational context, the utility of the helpfulness-based summarization approach is influenced by student prior teaching experience as well as student writing performance. Though low-performance students consider the baseline is the best for precision, students who have no teaching experience think HelpfulFilter is better than the baseline for content recall. More importantly, in the opinion of students who have teaching experience (more like domain experts), HelpfulSum is preferred over HelpfulFilter when comparing their content evaluation results on recall and accuracy.

In the future, we would like to build a fully automated summarizer by replacing the review helpfulness gold-standard with automated predictions as the filtering criteria. Given the collected human summaries, we will experiment with different feature combinations for sentence scoring and we will compare our helpfulness features with other content features as well. For summarizing peer reviews, the choice of review helpfulness gold-standard might also matter. A similar HelpfulSum but using expert-helpfulness ratings may generate most useful summaries, which would be an interesting follow-up study of the same topic. In addition, it seems that the traditional summarization method and the helpfulness-guided method capture useful information in different perspectives; using either one alone seems to be a tradeoff between precision and recall. In the future, we would like to see if using the combination of the traditional features and the helpfulness features yields better peer-review summarizers. Finally, with respect to our findings regarding the importance of differentiating review content (internal vs. external) for review helpfulness prediction, we would like to bring the same idea for review summarization. As the utility of review content of different content types for review helpfulness prediction varies with the domain, we might want to adjust the weight on different content types during content selection in accordance with the application domain.

6.0 CONCLUSIONS AND FUTURE WORK

In this work, we explore review helpfulness prediction and exploit review helpfulness for review summarization. Following prior work (Kim et al., 2006), we model review helpfulness prediction as a ranking problem, which can be solved by supervised machine learning based on features derived from review text and review context. While existing work on review helpfulness prediction has been dedicated to particular review domains such as Amazon product reviews, in this research, we provide two solutions for predicting review helpfulness in general settings: one is by specialization, and the other is through generalization. We first explore the specialization approach, investigating how existing helpfulness prediction techniques proposed for well studied domains can be tailored to a newer domain. In particular, we pick educational peer reviews as the target domain because of the educational semantics of helpfulness specific to the peer review. Then, we switch to the generalization approach, examining review (textual) content based on linguistic cues and content types in the same way across domains. More specifically, we propose a general review helpfulness model using standard computational linguistic features to capture the language usages, content diversity and helpfulness related topics in review content of different types. While the whole new set of features predict review-level helpfulness well in distinct review domains, the helpfulness-related topics can be used for review helpfulness analysis within a review.

In addition, we propose a novel review summarization method which leverages review helpfulness at different levels of granularity. We develop two helpfulness guided summarizers based on a standard multi-document extractive summarization framework using user-provided helpfulness assessment as our helpfulness gold standards. One summarizer uses the helpfulness gold-standard to filter out unhelpful reviews; the other further uses it to derive helpfulness-related topics and sentence-level helpfulness features, which replace the traditional features provided by MEAD for summarization sentence scoring. In contrast with existing work on review summarization, we show that user-provided helpfulness assessment can help review summarization, which naturally adapts to users' point of interest across domains. While early work merely used it as a filtering criteria before the summarization process, we demonstrate how it can be also used to infer review helpfulness at the sentence level for content selection directly.

By developing the peer-review helpfulness model, we show that techniques used in predicting product review helpfulness can be effectively adapted to the domain of peer reviews, with minor modification to address the domain speciality. Although the generic features proposed for product reviews are significantly correlated with peer-review helpfulness except for the metadata features (e.g., paper rating statistics) derived from review context, their utility varies between different review domains. Furthermore, to capture the educational semantics of peer-review helpfulness, we propose peer-review specialized features motivated by prior research in educational peer reviews, all of which yield high correlations with the helpfulness ratings. Our machine learning experimental results verify that adding the specialized features to the generic feature set enhances peer-review prediction performance. In particular, for capturing review lexical semantics, we find that lexical categories are preferred over unigrams for our peer review corpus (History2008); replacing the unigram features with the lexical category features reduces over-fitting, which enables other features to be added to further enhance the performance.

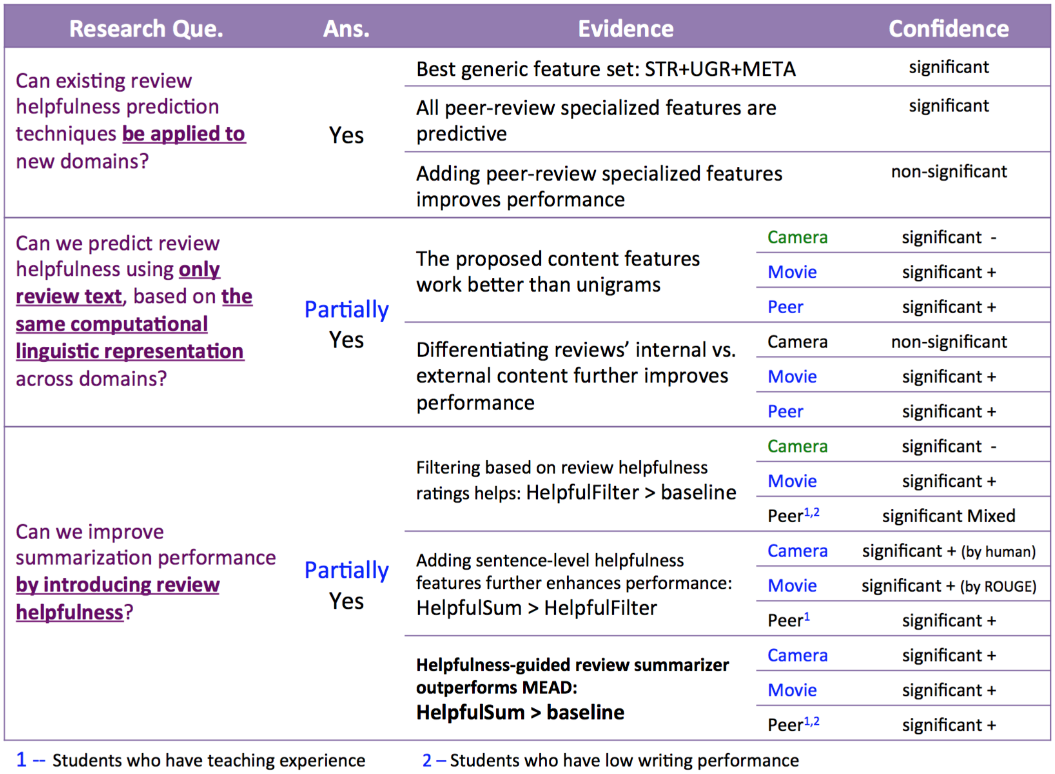
To avoid the domain expertise and human efforts required in the specialization approach, we alternatively propose a general and fully-automatic helpfulness model that can be applied to distinct review domains. In Chapter 4, we show that review helpfulness can be predicted by review language usage, content diversity and helpfulness related review topics using the same kind of feature representation in three distinct review domains (product reviews, movie reviews, and educational peer reviews). We provide a comprehensive analysis on the predictive power of each feature type in comparison with review unigrams in different feature combinations. Furthermore, we introduce the notion of review content categorization: separating a reviewer's evaluations and judgements (internal content) from the reviewer's references to the review subject (external content). To demonstrate the impact of the heterogeneity in review textual content on predicting review helpfulness, we experiment on the prediction task with the same feature extraction procedure, but varying the input text of feature extraction with respect to different content types. We show that different content (internal content vs. external content) have different predictive power, which also differs among review domains. However, performing content differentiation before feature extraction improves the model's helpfulness prediction performance for all three domains.

To demonstrate the value of review helpfulness for other review-related NLP tasks, we extend the scope of our research to review summarization, as a direct downstream application of review helpfulness prediction. In particular, we propose to use user-provided helpfulness assessment to identify useful review content for summarization purpose, because it naturally reflects users' point of interest in user interactions with online reviews. Although we are able to predict review helpfulness fully automatically (described above), in our summarization experiments, we use review helpfulness gold-standard to eliminate the helpfulness prediction noise. In addition to using review-level helpfulness for text pre-processing, we introduce the helpfulness-related topics used in our general helpfulness model, for developing sentence-level helpfulness features that can be used in extractive summarization algorithms directly. Our experiments based on MEAD show that our helpfulness guided summarizers are preferred over MEAD baseline for customer reviews, in which the summarizer that uses both the review-level and the sentence-level helpfulness information achieves the best performance in on both human and automated evaluation. When it comes to educational peer reviews, no significant difference between the summarizers is observed in human evaluation. Further analysis of student demographic background shows that the preference for helpfulness depends on student demographic background: students who received lower paper ratings are more likely to consider helpfulness-guided summarizers more helpful than the baseline; students who have no teaching experience are more likely to think the baseline more helpful than the helpfulness-guided ones. This makes us rethink about the validity of user-provided helpfulness assessment in the educational domain.

In terms of summarization evaluation, we consider both human evaluation and automatic evaluation metrics (ROUGE) for customer reviews. For both camera and movie reviews, we required participants to manually summarize the given reviews by selecting 10 sentences from them. Analysis on human summaries with respect to several effective heuristics proposed for news articles suggests that these heuristics cannot accurately reflect what most judges think useful, which provides empirical evidence for the need of developing new measurement of "importance" for user generated online reviews. As for the two human evaluation tasks, we find that pairwise comparison between summarizers yields more significant results than evaluating each summarizer in isolation. This is consistent with prior work (Lerman et al., 2009) which suggests that pairwise comparison is more suitable than evaluation in isolation for human evaluation.

To summarize, our research contributes to review helpfulness analysis in the perspectives of 1) automatic review helpfulness prediction and 2) utilizing review helpfulness for summarization. First, for prediction, we develop a peer-review helpfulness model – which demonstrates how to tackle the problem in a new review domain based on existing techniques developed for other domains, and a general helpfulness model – by exploring new computational linguistic features and differentiating review content in terms of internal content vs. external content. In addition to predicting helpfulness at the review level, we also investigate helpfulness prediction at the sentence level, based on helpfulness-related review topics learned through supervised LDA. Second, our work in review helpfulness prediction brings new ideas for extractive review summarization. In particular, we show that review helpfulness metadata (user provided review helpfulness assessment) can be used to generate sentence-level helpfulness features for summarization sentence scoring. Meanwhile, our work contributes to computer-supported online learning. For both helpfulness prediction and summarization, we explore a new domain – educational peer reviews, identifying new NLP challenges and providing solutions to address some of them. The proposed peer-review helpfulness model and our empirical findings in our peer-review summarization experiments will shed light on future work of building peer-review educational applications using AI and NLP. With respect to our hypotheses stated at the beginning of the thesis, we have summarized our main findings in Figure 6.1 in which the hypotheses are rephrased as research questions.

There are several remaining research questions that deserve consideration in the future. The first question is about the review helpfulness gold-standard. In this thesis, we use the percentage of "helpful" votes over all votes as the helpfulness gold-standard for customer reviews (movie reviews and peer reviews), though there are several limitations in it. First, different users may perceive the helpfulness of a review in different ways depending on the user needs. For example, different types of consumers may have different concerns (e.g., budget, fashion, functionality, etc.) when considering "whether I should buy this", and thus they seek different information in online reviews, which directly affects how they judge whether a review is helpful. User modeling and user adaptation may be used for predicting review helpfulness from the perspective of a particular user (group). Second, the helpfulness votes could be biased in several ways, such as the "early bird" effect (long existing reviews attract more readers and thus get more votes), "helpful reviews gets more helpful votes" (users are more likely to vote when they think a review is helpful), etc. One possible solution could be adjusting review helpfulness ratings mathematically based



Research Que.	Ans.	Evidence		Confidence
Can existing review helpfulness prediction techniques **be applied to** new domains?	Yes	Best generic feature set: STR+UGR+META		significant
		All peer-review specialized features are predictive		significant
		Adding peer-review specialized features improves performance		non-significant
Can we predict review helpfulness using **only review text**, based on **the same computational linguistic representation** across domains?	Partially Yes	The proposed content features work better than unigrams	Camera	significant -
			Movie	significant +
			Peer	significant +
		Differentiating reviews' internal vs. external content further improves performance	Camera	non-significant
			Movie	significant +
			Peer	significant +
Can we improve summarization performance **by introducing review helpfulness**?	Partially Yes	Filtering based on review helpfulness ratings helps: HelpfulFilter > baseline	Camera	significant -
			Movie	significant +
			Peer[1,2]	significant Mixed
		Adding sentence-level helpfulness features further enhances performance: HelpfulSum > HelpfulFilter	Camera	significant + (by human)
			Movie	significant + (by ROUGE)
			Peer[1]	significant +
		Helpfulness-guided review summarizer outperforms MEAD: HelpfulSum > baseline	Camera	significant +
			Movie	significant +
			Peer[1,2]	significant +

1 -- Students who have teaching experience   2 – Students who have low writing performance

Figure 6.1: A summary of the main findings.

on aligning the helpfulness ratings between reviews of similar textual content (Danescu-Niculescu-Mizil et al., 2009). With respect to educational peer reviews, it is important to capture the "true" helpfulness of peer reviews when choosing the gold-standard. In addition to expert helpfulness ratings, one alternative could be based on paper quality improvement if the paper revisions are available. Also, we would like to examine the impact of using different peer-review gold-standards for building our helpfulness-guided summarizers. We wonder if using expert helpfulness ratings can yield more useful summaries than traditional extractive summarizers such as MEAD.

In terms of feature engineering, as we have already built models to automatically predict certain cognitive constructs in educational peer reviews, we would like to use machine-generated codes instead of manual labels of these constructs so that we can further improve our peer-review specialized model in a fully automatic fashion. While we indirectly compare our two approaches of review helpfulness predictions (specialization vs. generalization) based on the peer review, we would like to compare them on camera review and movie review as well, by introducing existing helpfulness models of camera reviews and movie reviews respectively. For our generalization approach, in this thesis, we limit the size of our movie review corpus to be comparable to the peer reviews for evaluation concerns. For a comprehensive analysis of our helpfulness model on movie reviews, we would like to test it in larger scale. Also, we suspect that the number of movies included in the corpus may affect the helpfulness-related topics, especially for the external content. In the future, we would like to investigate how to control (or adapt to) this variation in our general helpfulness model. In a broader perspective, while we have examined review semantics (by means of various word-based features) and helpfulness-related review topics for predicting review helpfulness, we have not yet dealt with spelling errors and lexical ambiguity, which are common challenges in natural language understanding. Existing work in error correction and word sense disambiguation could be useful for our study of review helpfulness, which is another path to pursue for improving our work in the future.

As we have shown in the thesis, our work on review helpfulness prediction provides new opportunities for review summarization. Although we used helpfulness gold-standard as the review-level helpfulness ratings, in the future we would like to integrate our review helpfulness models into the helpfulness-guided summarizers to predict review-level helpfulness automatically. In the proposed helpfulness-guided summarizers, we incorporate sentence-level review helpfulness features for content selection, while we could also introduce differentiating review content for sentence scoring, in which the scoring function can weight different content types adaptively with respect to the review application domain.

Additionally, as it is suggested in our peer-review user study that traditional summarization method and our helpfulness-guided method capture useful information from different perspectives, we would like to see if a mixed method using both the traditional features and our helpfulness-features yields better summarizers, especially in the peer review domain. Also, we implemented our helpfulness-guided summarization method based on an extractive multi-document summarization framework, we wonder if this idea also works for other kinds of review summarizers, such as aspect-based opinion summarization systems. In a broader perspective, while we have only exploited review helpfulness for the summarization task, we believe that review helpfulness can be useful for other review-based applications as well. In terms of helpfulness prediction for user-generated content in general, the techniques proposed in this research on online reviews shed light on helpfulness analysis of other types of user generated content, such as online forum posts, user answers on social QA websites, etc.

APPENDIX

SUMMARIZATION USER STUDY MATERIALS

A.1 EXAMPLE OF CAMERA REVIEW SUMMARIES

This section provides an example of the **camera** review summarization test set, the corresponding summaries generated by the three summarizers, as well as human summaries.

A.1.1 Summarization test set

The following are 18 reviews on *Canon EF-S 60mm f2.8 Macro USM Lens for Canon SLR Cameras*, which is one of three summarization test sets that we used in our user study.

- This was one of the first lenses that I purchased with my Rebel XT. Now after taking several thousand pictures with it I can honestly say it was well worth the cost. I very rarely have any unsharp pictures with this lens unless it is my own fault by trying to use too slow of a shutter speed without a tripod.I also have the 180mm f3.5L Macro Lens, which costs about 3 times more than this lens, and although it is very clear and the extra reach is nice at times especially since it can be used with both the 1.4X and 2X TC's, it is very difficult to use inside without a tripod. The 60mm can be handheld if needed with very good results even if you have to bump your ISO up a little to do so.I have also used this lens for both inside and outside portrait work with very nice results.All-in-all, given it's small size and light weight I very rarely leave this lens behind when I go out because you never know when you might see a great macro shot.
- I love Canon products and I have had a complete Canon system for about 10 years. I love everything Canon does. However I do not understand Canon's reasoning behind producing this lens.The reason for the EF-S lenses is offering wider angle by getting

the rear elements of a lens closer to the “film” plane. They cannot do this on film and full frame sensor cameras because the mirror is larger and would hit the rear elements of an EF-S lens.They have indicated, however, that by 35 mm that advantage is gone. Why then do they produce a 60 mm Macro lens when they already have their macro requirements covered with they current three lenses? I would guess that the short back focus makes the lens cheaper; but this lens is only $60 short of the excellent 100 f2.8 USM Macro.As I said, if for $60 I can get a lens that has 40 mm more reach (66 in 1.6crop cameras; useful in macro photography) and that works on ALL CANON CAMERAS, I’m not going to be thinking about this lens at all.Granted, you might not be thinking of buying a FF camera anytime soon, but Canon has indicated that eventually they will have FF on all their DSLR’s, so why bother with this lens?The 17-85 or 10-22 are very reasonable offerings for the EF-S lineup, but the 60mm macro doesn’t make any sense to me.my $0.0.

- This lens is my favorite as I keep it on my Rebel XT at all times. I enjoy taking macro close-ups and portrait-type shots, which makes this a great lens for everyday use. It is light-weight and not bulky. My last SLR camera (years ago) was a Minolta with 50mm f/1.8 lens, and various other lenses that I rarely used. I usually don’t use a zoom lens due to the extra length and weight. Also, most non-professional zoom lenses are much slower at the closest tele-position due to the higher f/stop. With a fixed focal length of 60mm, I don’t mind moving myself toward or away from the subject (not a big deal). The pictures always appear to be sharp with good contrast and color saturation.My opinion on this lens is: “buy it... you’ll like it”. I did... and I love it. Good Luck.
- the optics are the best i’ve ever seen. the clarity is great. the abillity to focus on small objects only 2” away allows great macro pictures. however the auto focus is much slower than most canon lenses, but i can focus manually. this is my favotite lens ever
- This lens is top notch. The quality of the photographs is the best I have ever experienced with a camera.
- If you would like a list of sites with reviews email gumby at dontquotemeonthat dot com Pros: Very sharp, bright (fast), versatile, excellent build quality. Cons: AF tends to be dicy in low-light conditions. Pros: SHARP, SHARP, SHARP. No distortion, no CA, optically superb and better Cons: EF-S mount. Had to sell it when I upgraded to the 5D. Pros: Very solid build, Internal focusing is fantastic, Produces wonderfully saturated photos and high in contrast, Bokeh is lovely and very smooth, Auto Focus is typical ring-USM with Full Time Manual focus being excellent and smooth. f2.8 Aperture, 52mm Filter Size Cons: I do feel Canon could reduce the price, however for this quality I don’t mind paying for it. if your a newbie here’s some info A lens is “fast” when it has a low f-stop... ok so when you have a smaller number the apature is bigger which allows more light through, so this means you can up the shutter speed. and still have enough light reach the sensor. ok so lets say you have an out door shot if you have say an f/4 lens the shutter speed could be 1/250 of a second and you would get a good exposer. Now this lens can only go f/4 but if you in the same outdoor setting, had an f/2.8 lens you could jump to 1/500 of a second and get the same exposer. and freeze the action mmore effectivly, this i believe is why it’s a “fast” lens. ok have fun and get it done.
- The new Canon EF-S 60mm f2.8 Macro USM Digital SLR lens is designed to cover the entire field of the digital imaging sensors in Canon’s digital SLR line, most notably

the Canon EOS 20D. This corresponds in film to a normal lens perspective of approximately a 50mm lens. Furthermore it benefits from having Canon's USM autofocusing technology, allowing the photograher to have rapid, almost silent, autofocus, which is important when working in the field (You don't want to distract the animal you are photographing with the sound of the lens being focused.). Although this lens is not a L Series lens, the quality of its construction comes close to Canon's premium L Series professional line of lenses. Indeed, I have read elsewhere an excellent test report (I believe at Erwin Puts's website) on this lens praising its optical performance. Any Canon user of digital SLRs such as the EOS 20D who is interested in macro photography will regard this lens as absolutely essential for making great macro images.

- This lens is well worth the price. The first thing you will notice is the quality when you handle the lens. It has a very solid construction. But once you mount the lens is where it REALLY shines. The clarity of focus is the best I've seen and the focus is super fast. The macro functionality is just awesome. I highly recommend this lens.
- This lens is my favorite as I keep it on my Rebel XT at all times. I enjoy taking macro close-ups and portrait-type shots, which makes this a great lens for everyday use. It is light-weight and not bulky. My last SLR camera (years ago) was a Minolta with 50mm f/1.8 lens, and various other lenses that I rarely used. I usually don't use a zoom lens due to the extra length and weight. Also, most non-professional zoom lenses are much slower at the closest tele-position due to the higher f/stop. With a fixed focal length of 60mm, I don't mind moving myself toward or away from the subject (not a big deal). The pictures always appear to be sharp with good contrast and color saturation.My opinion on this lens is: "buy it... you'll like it". I did... and I love it. Good Luck.
- the optics are the best i've ever seen. the clarity is great. the abillity to focus on small objects only 2" away allows great macro pictures. however the auto focus is much slower than most canon lenses, but i can focus manually. this is my favotite lens ever.
- This lens is top notch. The quality of the photographs is the best I have ever experienced with a camera.
- If you would like a list of sites with reviews email gumby at dontquotemeonthat dot com Pros: Very sharp, bright (fast), versatile, excellent build quality. Cons: AF tends to be dicy in low-light conditions. Pros: SHARP, SHARP, SHARP. No distortion, no CA, optically superb and better Cons: EF-S mount. Had to sell it when I upgraded to the 5D. Pros: Very solid build, Internal focusing is fantastic, Produces wonderfully saturated photos and high in contrast, Bokeh is lovely and very smooth, Auto Focus is typical ring-USM with Full Time Manual focus being excellent and smooth. f2.8 Aperture, 52mm Filter Size Cons: I do feel Canon could reduce the price, however for this quality I don't mind paying for it. if your a newbie here's some info A lens is "fast" when it has a low f-stop... ok so when you have a smaller number the apature is bigger which allows more light through, so this means you can up the shutter speed. and still have enough light reach the sensor. ok so lets say you have an out door shot if you have say an f/4 lens the shutter speed could be 1/250 of a second and you would get a good exposer. Now this lens can only go f/4 but if you in the same outdoor setting, had an f/2.8 lens you could jump to 1/500 of a second and get the same exposer. and freeze the action mmore effectivly, this i believe is why it's a "fast" lens. ok have fun and get it done.
- This was one of the first lenses that I purchased with my Rebel XT. Now after taking

several thousand pictures with it I can honestly say it was well worth the cost. I very rarely have any unsharp pictures with this lens unless it is my own fault by trying to use too slow of a shutter speed without a tripod.I also have the 180mm f/3.5L Macro Lens, which costs about 3 times more than this lens, and although it is very clear and the extra reach is nice at times especially since it can be used with both the 1.4X and 2X TC's, it is very difficult to use inside without a tripod. The 60mm can be handheld if needed with very good results even if you have to bump your ISO up a little to do so.I have also used this lens for both inside and outside portrait work with very nice results.All-in-all, given it's small size and light weight I very rarely leave this lens behind when I go out because you never know when you might see a great macro shot.

- 4 starts because otherwise 5 stars is inevitable:1) Slow autofocus (hunt at times) but is to be expected of a macro lens.2) built quality not as solid as expected at this price range.I bought this lens instead for two purpose: Macro and Portrait! I was thinking of buying the 100mm macro plus 85mm f1.8 but this lens saves me buying two lenses! I have been very happy with it as what it is. I don't do flying insects very much so it is not a problem but that said I was able to get 1:1 shot of a fly, see sample here: http:www.theteh.comhtml3rd_350d_54.html There are other samples in this gallery here: http:www.theteh.comhtmlmy_3rd_350d_xt.html. The large aperture (F2.8) means that one could have shallow DOF and great for low light such as this pic: http:www.theteh.comhtml3rd_350d_49.html For portrait, I accidentally took this photo during the London Bombing of a women 'Shocked' by the incident unfolding in the public TV display. It was a coincident that her background inmage was the winning Reuter's photo of Tsunami tragedy and the matching colour of their dresses! I was quite far away so was able to capture her from head to toe: http:www.theteh.comhtml3rd_350d_25.html This illustrate the capability of both macro and normal photography using this lens. You will not regret it unless your primary aim is to shoot flying insets where longer 100mm or 150mm macro lenses may be needed in this case.
- I only wish that I had bought this lens earlier so I could have been using it longer. I am especially pleased with the short minimum focus distance - about 3 inches. This allows you to get very close to a small subject and to still fill the frame with the subject.I have had no problems with this lens and I love it.
- The new Canon EF-S 60mm f2.8 Macro USM Digital SLR lens is designed to cover the entire field of the digital imaging sensors in Canon's digital SLR line, most notably the Canon EOS 20D. This corresponds in film to a normal lens perspective of approximately a 50mm lens. Furthermore it benefits from having Canon's USM autofocusing technology, allowing the photograher to have rapid, almost silent, autofocus, which is important when working in the field (You don't want to distract the animal you are photographing with the sound of the lens being focused.). Although this lens is not a L Series lens, the quality of its construction comes close to Canon's premium L Series professional line of lenses. Indeed, I have read elsewhere an excellent test report (I believe at Erwin Puts's website) on this lens praising its optical performance. Any Canon user of digital SLRs such as the EOS 20D who is interested in macro photography will regard this lens as absolutely essential for making great macro images.
- This lens is well worth the price. The first thing you will notice is the quality when you handle the lens. It has a very solid construction. But once you mount the lens is

where it REALLY shines. The clarity of focus is the best I've seen and the focus is super fast. The macro functionality is just awesome. I highly recommend this lens.

- I love Canon products and I have had a complete Canon system for about 10 years. I love everything Canon does. However I do not understand Canon's reasoning behind producing this lens.The reason for the EF-S lenses is offering wider angle by getting the rear elements of a lens closer to the "film" plane. They cannot do this on film and full frame sensor cameras because the mirror is larger and would hit the rear elements of an EF-S lens.They have indicated, however, that by 35 mm that advantage is gone. Why then do they produce a 60 mm Macro lens when they already have their macro requirements covered with they current three lenses? I would guess that the short back focus makes the lens cheaper; but this lens is only $60 short of the excellent 100 f2.8 USM Macro.As I said, if for $60 I can get a lens that has 40 mm more reach (66 in 1.6crop cameras; useful in macro photography) and that works on ALL CANON CAMERAS, I'm not going to be thinking about this lens at all.Granted, you might not be thinking of buying a FF camera anytime soon, but Canon has indicated that eventually they will have FF on all their DSLR's, so why bother with this lens?The 17-85 or 10-22 are very reasonable offerings for the EF-S lineup, but the 60mm macro doesn't make any sense to me.my $0.0

A.1.2 Automatic summarization result

The summary generated by the baseline is:

1. All-in-all, given it's small size and light weight I very rarely leave this lens behind when I go out because you never know when you might see a great macro shot.
2. As I said, if for $ 60 I can get a lens that has 40 mm more reach (66 in 1.6 crop cameras; useful in macro photography) and that works on ALL CANON CAMERAS, I 'm not going to be thinking about this lens at all.
3. This lens is my favorite as I keep it on my Rebel XT at all times.
4. This lens is top notch.
5. I only wish that I had bought this lens earlier so I could have been using it longer.
6. The new Canon EF-S 60mm f2.8 Macro USM Digital SLR lens is designed to cover the entire field of the digital imaging sensors in Canon's digital SLR line, most notably the Canon EOS 20D.
7. Although this lens is not a L Series lens, the quality of its construction comes close to Canon's premium L Series professional line of lenses.
8. This lens is well worth the price.

The summary generated by HelpfulFilter is:

1. All-in-all, given it's small size and light weight I very rarely leave this lens behind when I go out because you never know when you might see a great macro shot
2. This lens is my favorite as I keep it on my Rebel XT at all times.

3. I only wish that I had bought this lens earlier so I could have been using it longer.
4. This lens is well worth the price.
5. The first thing you will notice is the quality when you handle the lens.
6. I also have the 180mm f/3.5 L Macro Lens, which costs about 3 times more than this lens, and although it is very clear and the extra reach is nice at times especially since it can be used with both the 1.4 X and 2X TC's, it is very difficult to use inside without a tripod.

The summary generated by HelpfulSum is:

1. Now after taking several thousand pictures with it I can honestly say it was well worth the cost.
2. I very rarely have any unsharp pictures with this lens unless it is my own fault by trying to use too slow of a shutter speed without a tripod.
3. I also have the 180mm f/3.5 L Macro Lens, which costs about 3 times more than this lens, and although it is very clear and the extra reach is nice at times especially since it can be used with both the 1.4 X and 2X TC's, it is very difficult to use inside without a tripod.
4. This was one of the first lenses that I purchased with my Rebel XT.
5. This lens is my favorite as I keep it on my Rebel XT at all times.
6. the abillity to focus on small objects only 2" away allows great macro pictures.
7. I enjoy taking macro close-ups and portrait-type shots, which makes this a great lens for everyday use.
8. the optics are the best i've ever seen.
9. 4starts because otherwise 5 stars is inevitable:
10. 1) Slow autofocus (hunt at times) but is to be expected of a macro lens.
11. I only wish that I had bought this lens earlier so I could have been using it longer.
12. This lens is well worth the price.

A.1.3 Human summary example

One participant selected the following ten sentences to form the summary:

1. Now after taking several thousand pictures with it I can honestly say it was well worth the cost.
2. This lens is my favorite as I keep it on my Rebel XT at all times.
3. The pictures always appear to be sharp with good contrast and color saturation.
4. The quality of the photographs is the best I have ever experienced with a camera.
5. Pros: SHARP, SHARP, SHARP. No distortion, no CA, optically superb and better Cons: EF-S mount. Had to sell it when I upgraded to the 5D.
6. Pros: Very solid build, Internal focusing is fantastic, Produces wonderfully saturated photos and high in contrast, Bokeh is lovely and very smooth, Auto Focus is typical ring-USM with Full Time Manual focus being excellent and smooth. f2.8 Aperture, 52mm Filter Size Cons: I do feel Canon could reduce the price, however for this quality I don't mind paying for it. if your a newbie here's some info A lens is "fast" when it has a low f-stop... ok

so when you have a smaller number the apature is bigger which allows more light through, so this means you can up the shutter speed.
7. this is my favotite lens ever.
8. the optics are the best i've ever seen.
9. 4 starts because otherwise 5 stars is inevitable
10. the clarity is great.

A.2 EXAMPLE OF MOVIE REVIEW SUMMARIES

This section provides an example of the **movie** review summarization test set and the corresponding summaries generated by the three summarizers.

A.2.1 Summarization test set

The following are 18 reviews on *The Lord of The Ring, The Return of The King*, which is one of three summarization test sets that we used in our user study for movie reviews.

- Thousands of comments have been made on this outstanding production and there is little left to write that has not already been written or said. Again, not surprisingly at last night's 'Oscars', the third film in the trilogy took most of the awards. Like others I could give glowing comments about content, acting, production, direction, visual effects, etc. but will instead, convey what I consider to be equally important; that is the realistic and accurate portrayal of a classic masterpiece of literature from one of the world's most imaginative authors. I have tried and failed three times to completely read the book and I enjoy reading, but feel that I could now do so and have a better understanding of the story - only because I know that Peter Jackson set out to retain accuracy of the story. Sometimes our own imagination lacks the ability to see exactly what the author intended and if a film can help that, then it only adds to the experience. By timely coincidence as I write this my computer screen saver has put up a picture of a mountain valley in New Zealand - it must know what is in my mind. That beautiful country was perhaps the ideal setting for the film with its mystical landscape punctuated with mountain valleys, rivers, forests and open spaces. It can not be far from what may have been in Tolkien's own mind.I would perhaps add one comment about content. Although there was much reliance on computer visualisation it was well-balanced by emotional acting like the characters Gollum and Gandalf. Although Gollum was a villain, I actually was made to feel sorry for him at the end. Too many potentially good films are spoilt by substituting acting for over indulgence in special effects. This is an art that the producers and directors of this film had exactly right.I hope that the success of this trilogy will herald a new era in film-making of classical stories. Our literature has a wealth of candidates, and even ones that have been tried could be re-visited now that such experiences as Lord of the Rings have proved financially viable and immensely popular.

- The Lord of the Rings: The Return of the King is, hands down, among the most spectacular and magnificent films of all time.A short run-down of the plot: After the battle of Helm's Deep and Saruman's imprisonment in his tower Orthanc, Aragorn, Legolas, Gimli and Gandalf re-group with Merry and Pippin in Isengard. There they learn that the army of Sauron is planning a full-scale attack on the largest city of men - Minas Tirith in Gondor. Gandalf and Pippin ride to Minas Tirith to warn Denethor, the steward of Gondor, of the threat from Mordor. Defenses are built up as the army of Sauron marches across the Pelennor Fields towards Minas Tirith. A distress call is sent to Rohan, still recovering from Helm's Deep. Rohan manage to muster a large army, and set out for Minas Tirith, but the battle has already begun. In the meantime, we continue with Sam and Frodo on their quest to destroy the One Ring.A major achievement of this epic film is the character development. Gollum becomes more cunning and sneaky than ever, and manages to turn Frodo against Sam, who is desperately trying his best to get his old Frodo back. Merry and Pippin are no longer just a source of comic relief, both of them prove themselves worthy as they are split up for the final battle. We learn about the true bravery and potential of hobbits as Merry helps cut down the Witch King. Eowyn also proves herself in the film, as she defies her uncle and sets out to Pelennor fields with the other Rohirrim, and eventually destroys the Witch King, and makes a very feminist remark while doing so. We learn to loathe Denethor because of his hatred of his last remaining son, Faramir, who really hasn't done anything wrong. The peak of our hatred for Denethor is reached in the scene where he tells Faramir that he would have preferred it if he had died instead of Boromir, his brother. And then, right after that, Denethor sends Faramir into certain suicide, and Faramir immediately accepts the mission he is appointed to, in a final attempt to please his father. And of course, Aragorn learns to accept his fate as the true king of men.In fact, the character development is so powerful that we actually participate in the character's feelings. We FEEL Frodo's exhaustion and agony as he literally drags himself across Mordor. We feel Sam's pain as Frodo is turned against him. And, just briefly, we participate in Gollum's triumph as he finally gets the One Ring. We are actually happy for Gollum and just for a brief moment, Frodo becomes the bad guy as he tries to take the ring back. All in all, Return of the King contains the most moving, emotional and touching scenes in the entire trilogy, and some of the best acting, especially from Sean Astin (Sam), Elijah Wood (Frodo), I an McKellen (Gandalf), John Noble (Denethor, he is very successful in adding depth to his character), Miranda Otto (Eowyn), and of course, Andy Serkis (Smeagol, and top-notch at it, just like in The Two Towers).The battle of Pelennor fields may be THE most spectacular and epic sequence in film history. Unlike Helm's Deep, Pelennor Fields shows the true cleverness of Sauron's army. Orcs are not the only participants; trolls are heavily used in the battle, as warriors and as beasts of burden. The nazgul are very significant in the battle, and while the Witch King didn't actually lead the battle as he did in the book, the nine ringwraiths and their fell beasts still play a key part and do lots of damage in the battle. We see just how powerful the nazgul really are. And of course, the men from the south and their massive oliphaunts play a significant part. While in Helm's Deep we felt triumphant, in Pelennor fields we only feel the triumph briefly, as the Rohirrim make their charge into the horde of orcs and trolls. The triumph in Pelennor Fields almost immediately dissolves, as the Rohirrim are trampled down by the oliphaunts. The battle is won, but we're not happy, we're grieved for all the destruction, all the losses. It's a totally different feeling than Helm's Deep, and makes this battle all the more superior.Return of the King features the most magnificent visuals in the entire trilogy. Whether they are of Minas Tirith, Pelennor Fields and Osgiliath, Mordor and the slopes of Mt. Doom or the climb to Shelob's cave near Minas Morgul, Peter Jackson really shows us the true impact of these landscapes and images.Many people may complain about the changes in the movie, especially the significant cut of Saruman from the end, but you must realize that if they would have featured the whole part with Saruman the movie would have continued another hour and a half. Don't fret; Peter Jackson said the scenes will all appear in the extended version of the film. The ending is long enough as it is, and the film continues at least another half an hour after the Ring is no more. The hobbits return to the shire, and Sam marries Rosie. Aragorn meets his fate and is crowned king, and is finally reunited with Arwen. And of course, one of the most moving scenes in the movie, in which Frodo gets on the last ship to the Undying lands with Bilbo, Gandalf, and the last of the elves (Galadriel and Elrond to name a few), and must part with his three hobbit friends for good.All in all, The Lord of the Rings: The Return of the King is one the most fine-tuned, cinematically perfect films ever made, it's absolutely flawless in every aspect. The Lord of the Rings trilogy as a whole is a spectacular achievement in film making history, and all three movies are together, without a doubt, the greatest epic ever made.

- It takes a miracle for me to go the cinema since smoking is banned in cinema chains but Peter Jackson is a miracle worker. How else would he be able to make me forget my filthy and disgusting nicotine addiction? He made me forget all about cigarettes for three hours with THE TWO TOWERS and I knew that with RETURN OF THE KING he could make me forget all about ciggies for a record breaking three and a half hours. I booked my ticket for Rothesay winter gardens cinema and sat down to be enthralled!!!!! SPOILERS!!!!! I do conclude there are some people in the world who can't see what the fuss is about with the LORD OF THE RINGS trilogy. My parents seem slightly puzzled that their cynical critical son loves LOTR. It's simply explained, these epic movies aren't a childish fantasy, they're like David Lean filming a Shakespeare play, but I do take onboard the criticism that the story structure of the movies can be irritating. FELLOWSHIP is very stop-start while the action intercutting in TTT can be annoying but ROTK has probably the best pace and structure of the three. ROTK starts with a sequence showing Smeagol murdering his friend in order to get the ring. This gives some needed backstory to Gollum. It also sets up its stall that it's not a family film never mind a "childish fantasy". In fact I predict that many of the children in Rothesay cinema will be having nightmares tonight due to the scenes with that horrible big spider, it made my skin crawl and the woman sitting next to me was gasping out loud as it prepared to cocoon poor Frodo, you should have seen the Q for the toilet after that scene which tells you how convincing the FX are in this movie, nothing appears CGI: Gollum isn't computer generated he's a living being and Peter Jackson doesn't use camera trickery for fight scenes he uses million upon millions of extras. He is David Lean reincarnated. No he is David Lean AND Will Shakespeare reincarnated, look at the way the cast act their parts, it's like they're appearing in the greatest play of the bard. Their performances are superlativeThere are some flaws. I did mention the script gives background to Smeagol but the script - Like the other films in the trilogy - is somewhat uneven. John Noble's character Denethor seems somewhat underwritten and I wasn't sure what his motivation was. Also as everyone else has mentioned the false endings are very irritating. When Aragorn is crowned king and the screen faded to black the audience reached for their Jackets and bags then we're shown another scene lasting several minutes which faded to a blaze of music. Everyone reached for their bags and jackets, then another scene which... It would have been better to have seen Aragorn crowned King and then seen Frodo sailing into the distance but I guess after the screenwriters have irritated us with the abrupt endings of the first two movies it's somewhat traditional to irritate with the end of the trilogy. These faults I can forgive but there is an unforgivable cop out of having an army of the undead charging to save the race of men from the Orc army at the end. It didn't ruin the movie for me but it just seems so lazy and contrived which stopped me from thinking it was the best movie in the trilogy, it's not, FELLOWSHIP is. But still this is a masterpiece of cinema which like cigarettes left me breathless and satisfied and hopefully we'll see it sweep the Oscar ceremony at lastAs for the Oscars themselves I'm puzzled about a couple of things. Howard Shore's score is beautiful and haunting but it's far from original with much of the music in ROTK re-used from FELLOWSHIP (The Gondor theme) and TTT (The Celtic music) while the omissions are even more surprising. No nominations for any of the actors! I know that all the great performances would cancel each other out but it's shameful Andy Serkis wasn't nominated as best supporting actor. Can anyone name a more unlikable baddie than Gollum in recent cinema? Me neither and no nomination for cinematography! I've no idea how John Lesnie's camera was able to keep up with the action and he deserved at least a nomination so maybe we'll see the third instalment robbed on Oscar night like FELLOWSHIP was. Even if it is that doesn't stop me and millions of other film fans from recognising the genuis of Peter Jackson. I bow to you Sir

- Fellowship of the Ring was far and away the best of the three Lord of the Rings movies, and the Academy snubbed it. The Two Towers was far less impressive, but that was understandable since the book of the Two Towers is the weakest of the original trilogy, and Jackson saved one of its best episodes, the confrontation between the hobbits and Shelob, for the third film. The third film rebounds, as it ought to have given that the third book is the best, but it does not reach the level reached by the first movie, much less by the book. Overall, Jackson did a good job, none of the movies is bad, and he deserves recognition for his work and the risks he took. It's just hard not to feel disappointed, given the huge promise of the first movie, to find that the trilogy as a whole is quite good but nowhere near great.Certainly Jackson achieved a very impressive feat in constructing battle scenes that are even more exciting and terrifying than the excellent ones in the previous two movies. The assault of Grond on the gate of Minas Tirith, the wild charge of the Rohirrim, the confrontation between Eowyn and the Lord of the Nazgul, and the desperate clash with the Oliphaunts are probably the finest fantasy warfare sequences ever filmed, managing to be intimate and detailed while also giving a sense of the overall strategic picture of the battle. Kurosawa would have been hard put to do better.Too, Jackson pulled a major coup by constructing a version of the climactic scene at Mount Doom that will surprise the readers of the original book without disappointing them; and it would have been very easy to go wrong at this point. And, Jackson manages a few times to do what he did with astonishing regularity in The Fellowship of the Ring: spot the dramatic moments and give them even more impact on film than they have on the printed page. His version of the scenes in the Paths of the Dead and the lighting of the beacons of Gondor are masterful.But, Jackson has lost his eye for character; indeed, he has lost it so disastrously that I have to wonder whether his master portraits of Boromir and Gandalf in the first film were anything more than luck. This is clearest in his revolting representation of Denethor. Jackson's Denethor is a cretin: weak, craven, stupid, self-pitying, insensitive, spiteful, utterly devoid of redeeming features. No man cut from this cloth could have lasted a month as Steward of Gondor, much less raised two of the boldest warriors of Minas Tirith or pitted his will against the Dark Lord Sauron for control of a Palantir. The true story of Denethor, which Jackson misunderstood completely, is not of the crumbling of a coward, but what is infinitely more tragic, the crumbling of a brave man.Meanwhile, Gandalf has receded into Old Testament prophet mode, and seems to have no emotions of his own whatsoever. Granted, even in the books Gandalf seems more distant and unapproachable after his reappearance, but he still had the old irritability and humor underneath. Arwen, after being used so well in the first movie, again becomes an annoying hindrance to the plot. Gimli, at least, has improved somewhat since The Two Towers; he is still being used as comic relief, but the humor is now more of a deliberately self-deprecating kind than the humiliating pratfall jokes he had to suffer through last time.Also, I have to complain about some of the things that Jackson left out. I will concede that he was right to omit two of my favorite parts: the meeting with Ghan-buri-Ghan and the Scouring of the Shire; time was limited, and something had to be cut. (he could have omitted the Paths of the Dead too, if he'd had to, although that would have been a shame considering how well he did that sequence). But the confrontation between Gandalf and the Witch-King of Angmar at the ruins of the Gate could have been done in thirty seconds, and the parley with the Mouth of Sauron would have required less than one minute to deliver one of the dramatic high points of the whole book.That Minas Tirith, Mount Doom, and the Grey Havens are magnificently done almost goes without saying. Art direction has been the one consistent strong point throughout this whole trilogy.In all, The Return of the King is a good movie. Certainly far worse ones have won Oscars. I just hope that the award doesn't lead to people imagining that this is the best movie of the trilogy.Rating: *** out of ****.Recommendation: Go see it on a big screen. But watch The Fellowship of the Ring first.

- In Return of the King - which follows the book (that I have not read, though heard what is in it that is not in the film) as close if not closer than the past two - co-writer/co-producer/director Peter Jackson brings Tolkien's grand tale of the quest to destroy the ring to an end. The story strands follow along the similar linear paths of the others, and it is done so with an equal worth in entertainment. Frodo, Sam and Gollum's path to Mordor unfolds as almost something of a love triangle for the ring; Merry and Pippen follow their own tales towards the great battle; Gandalf, Aragorn, Legolas, Gimli, and all the dwellers of middle earth prepare for the swarm of the terrors of Sauron. There is much praise that should be given to Jackson and his crew/cast on not just the worth of Return of the King, but to what is now the entire saga of the Lord of the Rings as a whole. Though the film does carry quite a load to it (at three hours and twenty-one minutes it's the longest of the three in theatrical form, and it definitely does go on at least ten to fifteen minutes longer than it should), and expands and deflates on the details of some characters (i.e. Saruman is nowhere in sight in this version, while Arwen gets more than what is from the original work), there are plenty of rousing scenes and sequences, terrific battles, and a grasp on the visual effects as a whole that don't let up. In all, ROTK is on the level with Fellowship and Two Towers, making the parts as good as the whole. This is something that only several other filmmakers can make a claim to, that one film does not bring on a let down from the expectations that preceded it. It's the kind of film I 'll want to see again, however it would be very difficult to sit through it in one place. Grade: A (both as a picture in and of itself, and overall on the three epics combined)

- Peter Jackson has done it. He has created an all-encompassing epic saga of Tolkien's Lord of the Rings books, and after coming away from the final chapter, how does this rate not only as a film on its own, but as a part of the whole? Perfect.I've never seen a series like this. A trilogy of movies created with such love and care and utter perfection of craft that you can't help but walk away and wonder how did Peter Jackson make this possible? I have always loved the original "Star Wars" and "Indiana Jones" series for their epic storytelling, and just for just fitting in as a great moment in cinema. This should be, will be, remembered with as much revered fondness for generations to come. They do not make films like these anymore.As a stand alone film, it picks up immediately where "Two Towers" ends, so brush up before seeing it. I've read the books, and the anticipation of seeing some of the more profound moments in this film made me kind of view it with a rushed sense of perspective. I wanted to make sure everything in this film was done "right". And when it happened, it was. I will need to see this again to enjoy everything on a more casual level.The cast comes through once more. The musical score retains its beauty, elegance and power. The special effects, notably Gollum again, are nothing less than breathtaking, and simply move the story along. The battles are monumentally huge and exciting. There are some liberties taken with the story, especially during the end with the homecoming, and yet, everything that needed to be covered regarding the main characters was handled. After the greatest moment of the series resolves itself, the story provided a breather. And gives a good-bye to friends seen on screen for the last three years. It was truly a bittersweet feeling in realizing that there will be no "Rings" movie in 2004. I will miss this talented group of actors.As with the first two, the film is very long, but goes by without you ever truly realizing it. This film is so much more than a simple "fantasy" epic. It's a story about strength of character, friendship, loyalty and love. And while every member of the Fellowship has their part to play, I finally understood why some critics have said this series is a story about Sam. It's his unwavering resolve that led the quest to its victory. Sean Astin is a true credit for adding the inspirational heart to this epic. As as far as the ending goes, they ended it the way that it had to be ended. Jackson ended this film the way it should have been.I will miss looking forward to a new "Rings" movie, but these movies provide hope that high-quality films can still be made without special effects taking over a story, bathroom humor, or a "Top 40" soundtrack. George Lucas could learn a lot from these films about how not to alienate the fanbase.Each film has earned a "10" from me for the last two years, which for me to give is a rarity. This one, however, is as equally deserving as its two predecessors. The Academy had better not look over this film for "Best Picture" of 2003. To do so would be greatly disrespectful of the craft and care that anyone involved with these films put into them.

- I think that almost everything that can be said about this trilogy has been said already, but still I will try. There are so many films that destroyed the beauty and perfection of the novels they have been built upon, not this one. In front of an amazingly beautiful scenery, Peter Jackson was able to create a fantasy-movie, which unlike so many others before did not deal with old clich & eacute; s and thus is far away from any trash-movie a lot of people had expected it to be beforehand.Although I am sure that the cast of this film will soon be forgotten, The "The Lord of the Rings" -trilogy will stand the times and be one of the most renowned pictures of the las decade.

- "The Lord of the Rings: Return of the King" is the third and final installment of Peter Jackson's adaptations of Tolkien's famous fantasy novels. Once again the makers of the film have taken care with the costumes, sets, scenery, models, CGI effects and Howard Shore's epic score to create a convincing depiction of Middle Earth.Once again the cast delivers expert performances. John Noble joins the cast as Denethor and effectively makes him into a despicable and repugnant character. Three of the performances in the film were particularly memorable for me. Bernard Hill once again brings authority to the role of King Theoden and his inspiring presence on the battlefield left me in awe. Miranda Otto brings strength to the role of Eowyn and makes the character's best moments unforgettable. I an McKellen once again brought his commanding presence as Gandalf to bear as he tried desperately to hold everything together.This film follows the familiar format of the first two films in taking Tolkien's work and streamlining it to create a well-paced film. The famous battle at Minas Tirith is on an unprecedented scale and the best fantasy battle ever filmed. As with the first two films, I found the added scenes for the extended addition interesting, but they didn't add much above and beyond the already great theatre cut.

- It's REALLY good. Every single thing about this movie is cool. It's my number one favourite movie of all time. (Well actually, the entire TRILOGY together is my favourite number one movie of all time.) There's no swearing or nudity. I still don't recommend it for the younger audience because there are some slightly frightening scenes, though. But anybody over eleven shouldn't be bothered. I don't recommend it for arachnophobia, because it might give them a heart attack. Anyways, this movie has an excellent beginning and a wonderful ending. And everything in the middle is great, too. BY ALL MEANS RENT IT, but make sure you watch the two first movies first.

- Like with the first two LotR movies, I hadn't (and still haven't I have to admit) read the books. So if your're looking for any comparison between the book and the movie, you have to look for other comments (and there are plenty of them here). The only thing I know, is that a small part of the end of book two (Two Towers) is implemented in Return of the King. Unfortunately for me, a review I read about the movie, did spoil that fact for me.In case you aren't aware of the book, I won't spoil anything that happens in this movie. But I'm going to assume that you have watched the previous installments ("Two Towers" and "Fellowship"). Our group has split up in 3 different smaller groups, each of which has it's own journey to go through. While Frodo and Sam have found themselves a companion in Gollum, the end of part two has hinted something dark that might happen here.That's one of the best things of the movies: The interaction and sometimes even seamless transitions between characters feelings for each other (Legolas vs. Gimli to name but one). Most of the time it's done in a subtle manner and even if it's played theatrically (here in this movie, a relationship between a father and his sons), it's impossible to defy the magic this movie brings onto your screen.The action is great (although a villain we can hang ourselves onto, like the leader of the Uruk-Hai in the first movie, is still missing) and the landscape phenomenal as in every movie of the series. While it was clear, that the actors wouldn't be considered Oscar material, they all bring their A-game and especially Viggo Mortensen is a revelation (makes you still wonder, what would've happened if the original actor that was cast for that role, went through with it).The main problem I see here (and many others have stated that also), lies at the end... well "endings". The movie doesn't seem to know, when to stop. So while you think all is over, you get another set-piece... and another one... and another one. While this might seem like nit-picking to some, some others were bothered very much by that fact. Still this is the best Fantasy Franchise that has hit our screens and makes other efforts seem pretty dull (Dungeons; Dragons anyone?). Now let's see if I can manage to read "The Hobbit" before they make a movie out of it... (I've already read a few pages)

- Before Peter Jackson's adaptation of The Lord of the Rings trilogy, the world of high fantasy has not been particularly well-served by cinema. The genre was not even really taken seriously in literature until the 1960s. During the 80s, there was a fad for fantasy movies, but while most of these looked nice and were good enough fun, none of them really had magnificence (although the 1981 Excalibur movie comes pretty close). It was not until the first decade of the 20th century that we saw fantasy cinema's rather delayed coming-of-age. As with the first two movies in the trilogy the transition from novel to screenplay is exceptional. There's a lot more action and a lot less dialogue in this one, and yet the plot is still clear and the narrative never feels repetitive. The idea of binding the various story lines together in time; such as when the Witch-King arises near Frodo and Sam, but the tower of green light is seen miles away by Pippin are great for building up the tension. They also really help to establish this vision of Middle Earth as a real place with vast dimensions.And again Jackson proves himself to be an action director with that little extra flair of intelligence. At first glance his work seems very much aimed at those with short attention spans, but there is so much loaded into each and every shot, the camera following an orc as he falls to the ground, or coming to rest upon a woman holding a baby as panic erupts in the city. His horror-tinged imagining of certain scenes is truly unnerving.There is some all-round improved acting in this instalment. Perhaps the years wrapped up in the production were taking the necessary toll on the cast. There are some truly heartfelt moments from Bernard Hill and a wonderfully spirited turn from Miranda Otto. For me, Billy Boyd always stood out as the finest of the hobbit performers, and this is the movie where he comes to the forefront, demonstrating great dignity and emotion. The best performance however, as previously, belongs to I an McKellen as Gandalf. There's something strangely knowing in his final scene.One of the unfortunate things about The Return of the King is that it suffers worse than the first two movies from a lack of dignity at certain times. The CGI Gollum is too cutesy and it's hard to believe in him as an antagonist, although funnily enough the glimpse we get of partly-transformed Smeagol biting into a fish with Andy Serkis in prosthetics would have been perfect for the whole thing. Some of the most serious bits become silly. I remember laughing out loud in the cinema when Gandalf says "So passes Denethor" when the man is still pathetically running around in flames.But by-and-large, this is an exceptional production, with its most outstanding touches in the way the whole thing has been put together. When the beacons are lit stretching a line across a mountain range, it's done in such a smooth, rhythmic way we are simultaneously impressed by the immense scale, the beauty of the landscape and the sheer brilliance of it as a means of communication. When Pippin's haunting song continues in the background as the men of Gondor ride off to their doom, we feel the depth of what is going on in a way the images alone could not impart. This is the kind of thinking you don't see in those numerous 80s fantasy movies, or in sci-fi's big trilogy, Star Wars. The Lord of the Rings movies put us right within both the excitement and the sadness of the story, for me with greater weight than Tolkien himself achieved. It elevates this above being merely another CGI action flick and grants the fantasy genre a status and stature it has never enjoyed before.

- WARNING: I advise anyone who has not seen the film yet to not read this comment.To tell the truth, I was actually very sad when I finished this film because it meant the fun of the Lord of the Rings series was over. The Lord of the Rings: The Return of the King was no doubt about it the best movie in the series and an excellent way to end a wonderful, captivating adventure.The acting from all three was A+, the suspense was A+, the battle sequences A+, and everything else was practically perfect in a sense, and I enjoyed it from beginning to end because of how well done it was. I loved how this movie makes you actually notice how all the characters have roughly changed through out their journey and how most of them's true colors show in this final film. Plus, the whole thing about Frodo writing a book and letting Sam be able to finish the last pages was an awesome script choice. I also loved how Smeagol takes a turn for the worse and battles it out with Frodo for the ring.Everything about The Lord of the Rings series is absolutely wonderful and this movie is incredibly incredible. I watched all three in one day, when I wasn't even expecting to like the first one, and now I consider them superb and well worth all the acclaim they receive. They'll remain in my mind for years to come, and I plan on watching them many more times in my life.

I can't believe I 'm saying this, but I love the Lord of the Rings!

- The hobbits approach the slopes of Mount Doom, preparing to dispose of the cursed Ring, while the forces of good and evil are rallied in anticipation of the ultimate battle. The film won the Academy Award for Best Picture, the only time in history a fantasy film has done so.The Return Of The King is the longest of the three films, which suffers from having to cut between disparate story strands, and - in its final half-hour - stacks up endings one after the other, like jet planes waiting to land, the director visibly reluctant to let these characters go. Most audiences will forgive Peter Jackson for that, for this is a fitting conclusion to a series of films made with tremendous artistry and affection for their subject; thrilling spectacle is underscored with palpable human drama, and it finally becomes clear why J.R.R. Tolkien's books continue to ring such bells so loudly in the lives of so many.

- Lord of the Rings: The Return of the King is no doubt the best movie I've seen. The film captures you instantly up to the words "The End" appearing on the screen. The Return of the King is nothing short of excitement.For all those who doesn't know what The Lord of the Rings is about (I'm thinking everybody does either through the books or the movies), here's a brief summary: Just over 50 year ago, J.R.R. Tolkien published a tale about a long, dangerous quest to destroy 1 ring. This tale is split into 3 novels, "The Fellowship of the Rings", "The Two Towers" and "The Return of the King". This tale, is called "The Lord of the Rings". Nearly 50 years after the novels were published, Peter Jackson tells the tale in another way, by bringing it to life onto the big screen.This tale is about 9 people's quest to destroy one magic ring and return peace to Middle-Earth. Ifthis one ring goes back to it's maker, the evil Sauron, the world will be under his control, bringing death and misery everywhere. The only way to destroy this powerful ring is to cast it into the fires of Mount Doom, where it was made. Only one hobbit could resist the temptation to keep the ring and this job was given to a hobbit named Frodo Baggins. To help him along the way, a fellowship of nine people were brought together. In it was a wizard, 2 men, an elf, a dwarf and three other hobbits. Together, they set out for Mordor, where Mount Doom is situated. Getting to Mordor is not as easy as it sounds. On the way they'll have to battle Orcs, Uruk-hais, giant spiders and other results of Tolkien's fascinating imagination.The Return of the King is the last addition to this vast trilogy. Peter Jackson does a great job in interpreting Tolkien's thoughts and feelings into a film. Watching the movies is just like reading the books themselves. If it was only the movies, it wouldn't have made such an impact on the public. The music in all three films practically takes you on your very own journey and the acting from the actors was also very impressive. Ifanything was missing from these movies, they just wouldn't have been such a success. Peter Jackson has done this to perfection.Overall, this is a must-watch movie. Plenty of action and special effects, not to mention a very heart warming ending to the trilogy. I'll just say one more thing, don't criticise anything until you have saw, heard or done it, especially Lord of the Rings.

- After a brief prologue showing us how agol came to get the ring and how he came to be the pitiful creature we know the film proper starts where The Two Towers left off;

agol is continuing with his plan to lure Frodo and Sam to their doom in Shelob's lair and Gandalf, Aragorn, Legolas and Gimli are reunited with Merry and Pippin at Isengard.When Pippin looks into the palantr, a sort of dark crystal ball, he sees the city of Minas Tirith under attack, unfortunately he himself is seen by Sauron. Galdalf takes him to Minas Tirith when he rides there to warn of the impending attack while the others remain behind to raise an army large enough to assist in that coming battle.I thought that the battle of Helm's Deep was impressive but it seems a mere skirmish compared to the battle of Minas Tirith, here there are thousands of Orcs supported by mercenaries mounted on gigantic elephant like creatures that are large enough to crush a man and his horse under their giant feet. All the time this battle rages Frodo and Sam are journeying deeper and deeper into the land of Mordor to Mount Doom in order to end Sauron's reign once and for all.This is probably the best film of the trilogy, I can see why it raked in the awards, the only weakness was the epilogue once the quest was completed the film could have ended at the crowning of the king rather than going on to their return to the Shire... even though I 'm aware that this was in the book. Peter Jackson did a fantastic job bringing such an epic story to the screen, he was of course assisted by a fantastic cast who really brought the characters to life.

- Feeling weary and battle-worn, I have just staggered out of the cinema after three and a half hours of special effects creatures fighting other special effects creatures. I had taken refreshments but barely touched them - probably because the film I had watched is one of the most mesmerising, evocative, inspiring, and awesome I have witnessed of any big adventure epic. Not to mention superb ensemble acting, moods that shift effortlessly between mediaeval battles of colossal proportions and convincing bloodshed, beauty and wonderment, fantastic natural and artificial landscapes and cityscapes, touches of humour, well-paced dramatic tension, and human bonding that is moving enough to just let you dry your eyes as the unassuming credits flash by.Return of the King is the greatest of the Tolkien trilogy by New Zealand director Peter Jackson. Although I've seen the other two and read the book, I felt it would also stand alone well enough for people who hadn't done either.The storytelling is much more professional that the first one - which maybe laboured to introduce so much information - or the second one - which has little let up from the tension of long battle scenes. In Return of the King, there is an emotional sting at the start, as we watch the transformation of Gollum from warm, fun-loving guy to murderous, mutated wretch. The movie then moves deftly between different segments of the story - the sadness of the lovely soft-focus Liv Tyler as fated Arwen whose travails and woman's love succeeds in having the Sword that was Broken mended, the comradeship of Sam and Frodo (Sean Astin Elijah Wood) that is tested to the limits, the strong commanding presence of Gandalf (I an McKellen) who keeps an eye on things whilst turning in an Oscar-worthy performance, the ingenious and very varied battle scenes, and the mythical cities of that rise out of the screen and provide key plot elements.This is a fairy story of human endeavour, the defeating of power cliques and the triumph of the human spirit that could almost be compared to Wagner's Gotterdammerung. It is a fairy story without any sugary sweetness, a fairy story the likes of which hasn't been told so well before, and is even unlikely to be done so well in the future. The haunting scream of the Nasgul stays with you, the physical attractions are not airbrushed, and the battles are about as far from pantomime characters waving wooden swords as you can get. The ingenious monsters keep you on the edge of your seat. The whole narrative maintains the spirit (if not archival, detailed accuracy) of the original and makes you want to read the book (or read the book again!) The worst I can say about it is that it is maybe a tad long - but not that you'd notice...

- This is the final movie in the Lord of the Rings trilogy, and certainly doesn't disappoint like some other trilogies *coughMatrixcough*. The three films had their principal shooting all done at the same time, which lowers their overall costs and keeps a good sense of continuity for the films.The special effects, first of all, are excellent. While there's a few little things (a reversed shot with smoke flowing back into chimneys and occasional lighting that's a bit off), by and large they're excellent. The most impressive thing about them is the sheer scale. This isn't a small or simple scene; it often includes thousands of digital characters combined with filmed actors and action, sweeping landscapes, and dozens of things happening at once. This is a good reason to see it in theatres; even on DVD, there's little details that you can only catch when it's on a massive screen.The filming is good, although there are a few evidences of digital smoothing and cutting that can nag at the mind and eyes of a picky movie-goer. There are a few interesting shots, but most are fairly plain and straight on, getting the point across without being dazzling. New Zealand's landscapes provide a great backdrop for everything going on, and there really are some beautiful places, especially up in the mountains. I hear land prices are quite good, what with the orcs warring and everything, so you may want to look into real estate purchases now.Sound has been said to make 75% of the emotional impact of any production. This is a loud 75%. All the sound effects are very well pulled off, sound appropriate, and are generally loud. The Nazgul screeching was bordering on painful, but in a good way. Most everything has a distinct sound, and it's rare that anything feels out of place. In some of the battles, the roof of the theatre was shaking. The soundtrack fits the movie well, and Howard Shore has done an excellent job, as with the last two films in the series.Performances all around were good, but Sean Astin as Sam and Viggo Mortensen as Aragorn really dominated the film. They performed their roles perfectly, and came away giving a good picture of the characters. Elijah Wood seemed to be stuck with the same terrified expression on his face through most of the movie, almost Max Payne-style, and it grew old quickly. I an McKellen, the ever-wise white wizard, had a fair bit of dialogue which he delivered well; my only complaint is he had too much in the way of wistful sayings leading to scene changes. Orlando Bloom, favorite of young teenage girls everywhere, had a few more action sequences (which got cheers from the aforementioned girls) which were quite well pulled off, but his acting wasn't much tested by this film. John Rhys-Davies continued with Gimli's joking performance; he's really too amusing to take seriously, but does a good job at it.For the old Tolkien fans, this movie stays quite close to the book, although they did have to omit some portions, most notably the taking and retaking of the Shire and the time spent in the Halls of Healing in Minas Tirith. Hopefully some of this will show up in the Extended Edition on DVD. Shelob's attack was left until this film, and much of the time spent in Mordor was shortened for the sake of pacing, and it was a good decision.My favorite scene would have to be the battle at Minas Tirith. The incredible scope of the battle, with the special effects, sounds, and many close-ups of pieces of the action, make for an exciting scene. The visual effects especially are stunning; the 'oliphaunts' play a big part in the action, and they're entirely created by computer. There's also some wide shots with tens of thousands of digital characters marching on the field of battle, and even the individual actions have the masses warring as a backdrop. It's worth your movie-going dollar simply to watch this on a large screen. It was also intermingled with some smaller events inside Minas Tirith, so it's not pure battle for the whole of the scene, and it keeps it from being dreary or heavy-handed.Overall, this is a movie well worth watching, and even paying to see in a theatre. I'd recommend against bringing small children, as there are some scary images, and they'd also be a distraction during the final movie in what will probably remain the series of the decade. Not a particularly great date movie, either... this is a real, bring-your-friends big movie. Five out of five decapitated orcs (and trust me, there were a lot more than that).

- ***SPOILERS*** ***SPOILERS***Over the years, I've read Lord of the Rings four times. During the holiday season of 20034, I watched Return of the King four times. While I embraced ROTK as the third part of a dream come true, I was not totally happy, left wondering why so many things vital were missing. The 4-hour extended DVD version explains a lot.My biggest beef was on so much missing about Aragon, and I found most of them in the DVD. One of the vital elements in the Fellowship's strategy is to draw Sauron's eye away from Frodo, and here Aragon's role is crucial. The "last debate" in the movie is totally inadequate in explaining the suicidal march to the Black gate but the DVD makes it very clear, with the additional scene of Aragon revealing himself to Sauron though the Palantir. He is the bait that Sauron can not resist.Another important aspect is that Aragon comes into the city of Minas Tirith first and foremost as a HEALER, not as a king. The kingship comes afterwards. This is again brought out in the additional scenes in the DVD, although still missing a lot of details from the book.Still disappointing, even for the DVD, is that so little is given to the story of Eowyn and Faramir. The dialogue through which they come to accept each other could very well be the most beautiful in the entire book. The few shots in the DVD that trace the development of their relationship are far from adequate, although that's a least a slight improvement from the film version.Another disappointment is Aragon's arrival at the Pelennor Fields, which is hopelessly lame compared with the original treatment in the book: amidst the despair of the Rohan and Gondor soldiers in witnessing the approaching black ships, Aragon's standard suddenly unfurls at the main mast: "There flowered a White Tree, and that was for Gondor; but seven stars were about it, and a high crown above it, the signs of Elendil that no lord had borne for years beyond count. And the stars flamed in the sunlight, for they were wrought of gems by Arwen daughter of Elrond; and the crown was bright in the morning, for it was wrought of mithril gold." The treatment of Gandalf's confrontation of the Witch King in the DVD departs from the book, in which the two are locked in a face off, then Rohan's horns are heard and the Witch King swings around and leaves. What in heaven's name is in Peter Jackson's mind when he had Gandalf's staff broken by the Witch King. But this did explain a mystery that has been bugging me for a year; why Gandalf had to snatch a spear from the guard when he saved Faramir from the pyre of Denethor.Enough on the DVD.I shall be remiss if I do not pay tribute to Peter Jackson for the wonderful film he and his dedicated crew have created.Most inspired is the lighting of the beacons to summon help from Rohan. In the book, this is observed by Pippin in the ride to Minas Tirith. To satisfy Pippin's curiosity, Gandalf explains the background to him in a somewhat factual manner. Jackson turns this into one of the most exciting moments in the film, with aesthetically superb shots of the 13 beacons (yes, I counted them) being lit up in succession, accompanied by beautifully rousing music score, culminating in Theoden's heroic utterance of "Rohan will answer". Watching this has to be among the most uplifting moments one can experience in a cinema.Most poignant is the Faramir's suicidal attempt to retake Osgilaith, under the orders of an unloving father. Starting from the soldiers of Gondor filing out of Minas Tirith in what looks almost like a funeral march to the letting loose of the swarm of arrows by the orcs in Osgilaith, every image of this scene is so hauntingly heartrending. It reminds me of John Woo's favourite scenes, although here, the music is Pipppin's actual singing rather than adapted background music, rendering the tragic mood even more devastating.Directly opposite in mood is Rohan's charge in the Battle of Pelennor Fields. Even if this mission is, in a way, equally suicidal, the spirit is sky high, radiating dauntless heroism and lust for battle. This scene also reminds me of the legendary battle scene in Spartacus (1960) which is universally recognised as the model in depiction of battle strategies. Rohan's charge in Pelennor Field, no the other hand, exemplifies heroism unsurpassed.Although ROTK is first and foremost the King's story, we should not forget, in the overall scheme of things, the ring bearers (no typo here because Frodo did acknowledge Sam as a fellow ring bearer in the end of the book). Elijah Wood and Sean Astin (particularly Astin) have played their roles to perfection. Towards the end of the quest, when Frodo's strength was almost fully spent, to hear Sam say "I can not carry it (the ring) for you, Mr. Frodo, but I can carry you" and not be moved, one will have to be a hopelessly and irreversibly hardened cynic. The background music, incidentally, is "Into the west".It is certainly a good sign that the general audience worldwide has reacted favourably to the long aftermath following the destruction of the ring, indicated that their capacity to appreciate has not been impaired by the proliferation of Hollywood style slam-bang endings. Viggo Mortensen's line to the Hobbits "My friends, you bow to no one" is delivered with sincerity and conviction. The final scene at the Grey Havens is graceful, touching, stylish. However, there is one shot that I must mention: Galadriel's final enigmatic, alluring, half-smiling glance at Frodo before she disappears into the ship. Cate Blanchett is among the most versatile actresses around today and in LOTR, she is Galadriel incarnate.

A.2.2 Automatic summarization result

The summary generated by the baseline is:

1. In fact I predict that many of the children in Rothesay cinema will be having nightmares tonight due to the scenes with that horrible big spider, it made my skin crawl and the woman sitting next to me was gasping out loud as it prepared to cocoon poor Frodo, you

should have seen the Q for the toilet after that scene which tells you how convincing the FX are in this movie, nothing appears CGI : Gollum isn't computer generated he's a living being and Peter Jackson doesn't use camera trickery for fight scenes he uses million upon millions of extras.
2. In Return of the King - which follows the book (that I have not read, though heard what is in it that is not in the film) as close if not closer than the past two - co-writer/co-producer/director Peter Jackson brings Tolkien's grand tale of the quest to destroy the ring to an end.
3. Lord of the Rings : The Return of the King is no doubt the best movie I've seen.
4. After a brief prologue showing us how Sm agol came to get the ring and how he came to be the pitiful creature we know the film proper starts where The Two Towers left off; agol is continuing with his plan to lure Frodo and Sam to their doom in Shelob's lair and Gandalf, Aragorn, Legolas and Gimli are reunited with Merry and Pippin at Isengard.

The summary generated by HelpfulFilter is:

1. In fact I predict that many of the children in Rothesay cinema will be having nightmares tonight due to the scenes with that horrible big spider, it made my skin crawl and the woman sitting next to me was gasping out loud as it prepared to cocoon poor Frodo, you should have seen the Q for the toilet after that scene which tells you how convincing the FX are in this movie, nothing appears CGI : Gollum isn't computer generated he's a living being and Peter Jackson doesn't use camera trickery for fight scenes he uses million upon millions of extras.
2. "The Lord of the Rings : Return of the King" is the third and final installment of Peter Jackson's adaptations of Tolkien's famous fantasy novels.
3. Like with the first two LotR movies, I hadn't (and still haven't I have to admit) read the books.
4. WARNING : I advise anyone who has not seen the film yet to not read this comment.
5. Lord of the Rings : The Return of the King is no doubt the best movie I've seen.

The summary generated by HelpfulSum is:

1. Peter Jackson has done it.
2. A short run-down of the plot : After the battle of Helm's Deep and Saruman's imprisonment in his tower Orthanc, Aragorn, Legolas, Gimli and Gandalf re-group with Merry and Pippin in Isengard.
3. I think that almost everything that can be said about this trilogy has been said already, but still I will try.
4. "The Lord of the Rings : Return of the King" is the third and final installment of Peter Jackson's adaptations of Tolkien's famous fantasy novels.
5. Like with the first two LotR movies, I hadn't (and still haven't I have to admit) read the books.
6. WARNING : I advise anyone who has not seen the film yet to not read this comment.
7. The hobbits approach the slopes of Mount Doom, preparing to dispose of the cursed Ring, while the forces of good and evil are rallied in anticipation of the ultimate battle.

8. Feeling weary and battle-worn, I have just staggered out of the cinema after three and a half hours of special effects creatures fighting other special effects creatures.

A.2.3 Human summary example

One participant selected the following ten sentences to form the summary:

1. Like others I could give glowing comments about content, acting, production, direction, visual effects etc. but will instead, convey what I consider to be equally important ; that is the realistic and accurate portrayal of a classic masterpiece of literature from one of the world's most imaginative authors.
2. I have tried and failed three times to completely read the book and I enjoy reading, but feel that I could now do so and have a better understanding of the story - only because I know that Peter Jackson set out to retain accuracy of the story.
3. The 4-hour extended DVD version explains a lot.
4. My biggest beef was on so much missing about Aragon, and I found most of them in the DVD.
5. Many people may complain about the changes in the movie, especially the significant cut of Saruman from the end, but you must realize that if they would have featured the whole part with Saruman the movie would have continued another hour and a half.
6. Peter Jackson said the scenes will all appear in the extended version of the film.
7. The third film rebounds, as it ought to have given that the third book is the best, but it does not reach the level reached by the first movie, much less by the book.
8. Overall, Jackson did a good job, none of the movies is bad, and he deserves recognition for his work and the risks he took.
9. It's just hard not to feel disappointed, given the huge promise of the first movie, to find that the trilogy as a whole is quite good but nowhere near great.
10. The Lord of the Rings : The Return of the King is, hands down, among the most spectacular and magnificent films of all time.

A.3 ONLINE SURVEY OF PEER REVIEWS

The survey questions that we asked student "isabella-aqua-3", who is a female English native speaker, with teaching experience. Her average paper rating is 5.4 (low performance), average review-helpfulness is 4.8 (high performance), and average review accuracy is .85 (high performance).

A.3.1 Introduction

At the beginning of the user study, we inform the students the purpose of the experiment as well as the general experimental setup:

> **Evaluation of Helpfulness-Guided Extractive Review Summarization**
> The purpose of this research study is to evaluate the effectiveness of our proposed framework for automatically summarizing peer reviews. For that reason, we are running this survey and ask SWoRD/ARROW users to assess the quality of machine-generated summaries of the peer reviews that you received (in your 1st assignment of PHYS 0212 Introduction to Laboratory Physics).
> In this survey, you will compare 3 pairs of summaries of the reviews that you received in the first assignment of PHYS0212. You will also rate each summary on three dimensions. This won't take long (15 mins), especially because you are familiar with the content and you have read the reviews already. Once you complete the survey (all required entries), you will get a five dollars amazon-gift card as a reward.
> Your participation is voluntary, and you may withdraw from this study at any time. This study is being conducted by Wenting Xiong, who can be reached by email wex12@cs.pitt.edu if you have any questions. We appreciate your participation to help our research.

A.3.2 Pairwise comparison question

Question 1. Here are two summaries (Table A.3.2) about the set of reviews you just read. Which one of them is more helpful/constructive?

1. Strongly prefer A 2. Slightly prefer A 3. no preference 4. Slightly prefer B 5. Strongly prefer B

Question 2. Here are two summaries (Table A.3.2) about the set of reviews you just read. Which one of them is more helpful/constructive?

1. Strongly prefer A 2. Slightly prefer A 3. no preference 4. Slightly prefer B 5. Strongly prefer B

Question 3. Here are two summaries (Table A.3.2) about the set of reviews you just read. Which one of them is more helpful/constructive?

1. Strongly prefer A 2. Slightly prefer A 3. no preference 4. Slightly prefer B 5. Strongly

Summary A	Summary B
1. The first sentence in the ohmmeter section appears to be missing a value following the word ‘of’, and any exclamation points should be removed. 2. The style of the conclusion seems to be much different than the rest of the paper. 3. The sentences are much more direct and have less punctuation, especially commas, which made it easier to read. 4. General spelling errors were present. 5. I noticed one at the bottom of page 5 (recoreded instead of recorded) and onther on page 4 (and instead of an). 6. Good use of parenthesis to explain. 7. Very consistent with the rest of the format you used. 8. I would move the title of “Data Analysis” down one line so it is above the body of the section for clarity on when the next section starts and fluidity purposes. 9. While well written, when the topics switch from ammeter, ohmmeter, or volmeter there is typically an abrupt ending followed by a sentence on the next device. 10. The abstract could be clearer in differentiating the parts of the experiment and individually reporting those numerical results from the data analysis. 11. Very concise and well written, gets straight to the point.	1. The first sentence in the ohmmeter section appears to be missing a value following the word ‘of’, and any exclamation points should be removed. 2. The style of the conclusion seems to be much different than the rest of the paper. 3. I noticed one at the bottom of page 5 (recoreded instead of recorded) and onther on page 4 (and instead of an). 4. Good use of parenthesis to explain. 5. The second paragraph of the introduction has a typo of saying the unit is “loules” rather than “Joules”. 6. I would move the title of “Data Analysis” down one line so it is above the body of the section for clarity on when the next section starts and fluidity purposes. 7. While well written, when the topics switch from ammeter, ohmmeter, or volmeter there is typically an abrupt ending followed by a sentence on the next device. 8. The abstract could be clearer in differentiating the parts of the experiment and individually reporting those numerical results from the data analysis. 9. Very concise and well written, gets straight to the point.

Table A1: Peer review survey example – pairwise comparison between HelpfulSum (left) and HelpfulFilter (right). Student rating = 2.

prefer B

Summary A	Summary B
1. The first sentence in the ohmmeter section appears to be missing a value following the word ‘of’, and any exclamation points should be removed. 2. The style of the conclusion seems to be much different than the rest of the paper. 3. I noticed one at the bottom of page 5 (recoreded instead of recorded) and onther on page 4 (and instead of an). 4. Good use of parenthesis to explain. 5. The second paragraph of the introduction has a typo of saying the unit is “loules” rather than “Joules”. 6. I would move the title of “Data Analysis” down one line so it is above the body of the section for clarity on when the next section starts and fluidity purposes. 7. While well written, when the topics switch from ammeter, ohmmeter, or volmeter there is typically an abrupt ending followed by a sentence on the next device. 8. The abstract could be clearer in differentiating the parts of the experiment and individually reporting those numerical results from the data analysis. 9. Very concise and well written, gets straight to the point.	1. Overall, this section was clearly explained and very concise. 2. In the internal resistance of the galvanometer section, the statement “which should have equaled R (m)” does not need to be included, as that should be discussed in data analysis. 3. While a part of your results, the “poor agreement” or “abysmal agreement” portions don’t seem necessary as part of the initial abstract in providing a summary of the report and could be cut out for a more appropriate length. 4. I would move the title of “Data Analysis” down one line so it is above the body of the section for clarity on when the next section starts and fluidity purposes. 5. Incorporating how the data connects to the theory of the experiment or connecting it to the introduction may aid in better understanding the values 6. Much of the information may be relevant but could be moved to the data analysis section near it’s corresponding data values. 7. The abstract could be clearer in differentiating the parts of the experiment and individually reporting those numerical results from the data analysis.

Table A2: Peer review survey example – pairwise comparison between HelpfulFilter (left) and the baseline (right). Student rating = 4.

A.3.3 Content evaluation questions

Question 4. Consider the following summary (Table A.3.3) only. How do you agree with the following statement regarding the summary content? (1: Strongly disagree 2: Disagree 3: Neither agree nor disagree 4: Agree 5: Strongly agree)

Summary A	Summary B
1. The first sentence in the ohmmeter section appears to be missing a value following the word ‘of’, and any exclamation points should be removed. 2. The style of the conclusion seems to be much different than the rest of the paper. 3. The sentences are much more direct and have less punctuation, especially commas, which made it easier to read. 4. General spelling errors were present. 5. I noticed one at the bottom of page 5 (recoreded instead of recorded) and onther on page 4 (and instead of an). 6. Good use of parenthesis to explain. 7. Very consistent with the rest of the format you used. 8. I would move the title of “Data Analysis” down one line so it is above the body of the section for clarity on when the next section starts and fluidity purposes. 9. While well written, when the topics switch from ammeter, ohmmeter, or volmeter there is typically an abrupt ending followed by a sentence on the next device. 10. The abstract could be clearer in differentiating the parts of the experiment and individually reporting those numerical results from the data analysis. 11. Very concise and well written, gets straight to the point.	1. Overall, this section was clearly explained and very concise. 2. In the internal resistance of the galvanometer section, the statement “which should have equaled R (m)” does not need to be included, as that should be discussed in data analysis. 3. While a part of your results, the “poor agreement” or “abysmal agreement” portions don’t seem necessary as part of the initial abstract in providing a summary of the report and could be cut out for a more appropriate length. 4. I would move the title of “Data Analysis” down one line so it is above the body of the section for clarity on when the next section starts and fluidity purposes. 5. Incorporating how the data connects to the theory of the experiment or connecting it to the introduction may aid in better understanding the values 6. Much of the information may be relevant but could be moved to the data analysis section near it’s corresponding data values. 7. The abstract could be clearer in differentiating the parts of the experiment and individually reporting those numerical results from the data analysis.

Table A3: Peer review survey example – comparison between HelpfulSum (left) and the baseline (right). Student rating = 4.

1. It covers ALL information I would like to include. (content recall)

2. It contains NO information that I would NOT have included. (content precision)

3. It reflects the ideas of reviews accurately. (content accuracy)

1. The first sentence in the ohmmeter section appears to be missing a value following the word ‘of’, and any exclamation points should be removed.
2. The style of the conclusion seems to be much different than the rest of the paper.
3. The sentences are much more direct and have less punctuation, especially commas, which made it easier to read.
4. General spelling errors were present.
5. I noticed one at the bottom of page 5 (recoreded instead of recorded) and onther on page 4 (and instead of an).
6. Good use of parenthesis to explain.
7. Very consistent with the rest of the format you used.
8. I would move the title of “Data Analysis” down one line so it is above the body of the section for clarity on when the next section starts and fluidity purposes.
9. While well written, when the topics switch from ammeter, ohmmeter, or volmeter there is typically an abrupt ending followed by a sentence on the next device.
10. The abstract could be clearer in differentiating the parts of the experiment and individually reporting those numerical results from the data analysis.
11. Very concise and well written, gets straight to the point.

Table A4: Peer review survey example – content evaluation on HelpfulSum. The student ratings: recall = 4, precision = 2, accuracy = 4.

Question 5. Consider the following summary only (Table A.3.3). How do you agree with the following statement regarding the summary content? (1: Strongly disagree 2: Disagree 3: Neither agree nor disagree 4: Agree 5: Strongly agree)

1. It covers ALL information I would like to include. (content recall)
2. It contains NO information that I would NOT have included. (content precision)
3. It reflects the ideas of reviews accurately. (content accuracy)

Question 6. Consider the following summary only (Table A.3.3). How do you agree with the following statement regarding the summary content? (1: Strongly disagree 2: Disagree 3: Neither agree nor disagree 4: Agree 5: Strongly agree)

1. It covers ALL information I would like to include. (content recall)
2. It contains NO information that I would NOT have included. (content precision)

1. Overall, this section was clearly explained and very concise.
2. In the internal resistance of the galvanometer section, the statement “which should have equaled R (m)” does not need to be included, as that should be discussed in data analysis.
3. While a part of your results, the “poor agreement” or “abysmal agreement” portions don’t seem necessary as part of the initial abstract in providing a summary of the report and could be cut out for a more appropriate length.
4. I would move the title of “Data Analysis” down one line so it is above the body of the section for clarity on when the next section starts and fluidity purposes.
5. Incorporating how the data connects to the theory of the experiment or connecting it to the introduction may aid in better understanding the values
6. Much of the information may be relevant but could be moved to the data analysis section near it’s corresponding data values.
7. The abstract could be clearer in differentiating the parts of the experiment and individually reporting those numerical results from the data analysis.

Table A5: Peer review survey example – content evaluation on the baseline. The student ratings: recall = 3, precision = 2, accuracy = 4.

3. It reflects the ideas of reviews accurately. (content accuracy)

A.4 PEER REVIEWS THAT THE STUDENT “ISABELLA-AQUA-3” RECEIVED

- You did a good job of including definitions and analogies at the beginning of the report. These are very helpful in making the theory clear.
- Overall, this section was clearly explained and very concise. In the internal resistance of the galvanometer section, the statement “which should have equaled R(m)” does not need to be included, as that should be discussed in data analysis.
- The first sentence in the ohmmeter section appears to be missing a value following the word ‘of’, and any exclamation points should be removed.
- The style of the conclusion seems to be much different than the rest of the paper. The sentences are much more direct and have less punctuation, especially commas, which made it easier to read. Some of the other sections have sentence structures that rely heavily on commas, and it makes it more difficult to read, in my opinion. Maybe try to replicate the conclusion’s style in the rest of the paper.
- General spelling errors were present. I noticed one at the bottom of page 5 (recoreded instead of recorded) and onther on page 4 (and instead of an).

1. The first sentence in the ohmmeter section appears to be missing a value following the word ‘of’, and any exclamation points should be removed.
2. The style of the conclusion seems to be much different than the rest of the paper.
3. I noticed one at the bottom of page 5 (recoreded instead of recorded) and onther on page 4 (and instead of an).
4. Good use of parenthesis to explain.
5. The second paragraph of the introduction has a typo of saying the unit is “loules” rather than “Joules”.
6. I would move the title of “Data Analysis” down one line so it is above the body of the section for clarity on when the next section starts and fluidity purposes.
7. While well written, when the topics switch from ammeter, ohmmeter, or volmeter there is typically an abrupt ending followed by a sentence on the next device.
8. The abstract could be clearer in differentiating the parts of the experiment and individually reporting those numerical results from the data analysis.
9. Very concise and well written, gets straight to the point.

Table A6: Peer review survey example – content evaluation on HelpfulFilter. The student ratings: recall = 4, precision = 4, accuracy = 4.

- Good use of parenthesis to explain.
- Could organize format of equations better. Instead of bolding Equation and number underneath the equation, put on the same line.
- Great explanation of theory with set up.
- Very consistent with the rest of the format you used.
- Targeted at the appropriate audience level.
- I did not see any major problems with grammar. I would probably change the font to something more common.
- While a part of your results, the “poor agreement” or “abysmal agreement” portions don’t seem necessary as part of the initial abstract in providing a summary of the report and could be cut out for a more appropriate length.
- The second paragraph of the introduction has a typo of saying the unit is “loules” rather than “Joules”.
- I would move the title of “Data Analysis” down one line so it is above the body of the section for clarity on when the next section starts and fluidity purposes.
- The presentations of values of collected data are accurated recorded and some further percent error was included with this information for understanding on its accuracy. Incorporating how the data connects to the theory of the experiment or connecting it to the introduction may aid in better understanding the values
- The conclusion, being over a page single spaced, must be shortened to concisely summarize the experiment. Much of the information may be relevant but could be moved to the data analysis section near it’s corresponding data values. For example, the

interpretation and percent error on the slope of the resistence and current plot could be relocated near Figure 4, which depics this data.

- While well written, when the topics switch from ammeter, ohmmeter, or volmeter there is typically an abrupt ending followed by a sentence on the next device. May want to include transitional sentences, or combine thoughts which incorporate multiple devices, at least for the abstract and conclusion portions of the lab.
- The abstract could be clearer in differentiating the parts of the experiment and individually reporting those numerical results from the data analysis.
- Purpose: To measure current, voltage, and resistance across various circuit setups. There is a good balance between theory and equations. The author does a good job explaining specific terminology.
- The experimental section provides enough details to reproduce the experiment. All of the figures are clearly explained.
- All of the tables and graphs are clearly labeled.
- The sources of error are well explained in the conclusion. Possible improvements or ways of re-evaluated the data are provided.
- "R2 was then recoreded as well"
- Very concise and well written, gets straight to the point.

BIBLIOGRAPHY

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 579–586. Association for Computational Linguistics, 2005.

Maya Ando and Shun Ishizaki. Analysis of travel review data from reader's point of view. *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 47–51, 2012.

Yigal Attali and Jill Burstein. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3), 2006.

Regina Barzilay and Lillian Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *HLT-NAACL*, pages 113–120, 2004.

Steven Bethard, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky. Automatic extraction of opinion propositions and their holders. In *2004 AAAI Spring Symposium on Exploring Attitude and Affect in Text*, page 2224, 2004.

David M Blei and Jon D McAuliffe. Supervised topic models. *arXiv preprint arXiv:1003.0783*, 2010.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

Samuel Brody and Noemie Elhadad. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812, 2010.

Qing Cao, Wenjing Duan, and Qiwei Gan. Exploring determinants of voting for the “helpfulness” of online user reviews: A text mining approach. *Decision Support Systems*, 50 (2):511–521, 2011.

Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336. ACM, 1998.

Giuseppe Carenini, Raymond T Ng, and Adam Pauls. Multi-document summarization of evaluative text. In *In Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

Kwangsu Cho. Machine classification of peer comments in physics. In *Proceedings of the First International Conference on Educational Data Mining (EDM2008)*, pages 192–196, 2008. URL http://www.educationaldatamining.org/EDM2008/uploads/proc/21_Cho_32.pdf.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 355–362. Association for Computational Linguistics, 2005.

Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinber g, and Lillian Lee. How opinions are received by online communities: A case study on Amazon .com helpfulness votes. In *Proceedings of the 18th International Conference on World Wide Web*, pages 141–150, 2009.

Ruihai Dong, Markus Schaal, Michael P O’Mahony, and Barry Smyth. Topic extraction from online reviews for classification and recommendation. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 1310–1316. AAAI Press, 2013.

Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.(JAIR)*, 22(1):457–479, 2004.

Andrea Ernst-Gerlach and Gregory Crane. Identifying quotations in reference works and primary materials. In *Research and Advanced Technology for Digital Libraries*, pages 78–87. Springer, 2008.

Raquel M. Crespo Garcia. Exploring document clustering techniques for personalized peer assessment in exploratory courses. In *Proceedings of Computer-Supported Peer Review* *in Education (CSPRED) Workshop in the Tenth International Conference on Intelligent Tutoring Systems*, 2010.

Anindya Ghose and Panagiotis G Ipeirotis. 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, 2011.

Dan Gillick and Benoit Favre. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, pages 10–18. Association for Computational Linguistics, 2009.

Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202, 2004.

Aria Haghighi and Lucy Vanderwende. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370. Association for Computational Linguistics, 2009.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

Jeff Huang, Oren Etzioni, Luke Zettlemoyer, Kevin Clark, and Christian Lee. Revminer: An extractive interface for navigating reviews on a smartphone. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, pages 3–12. ACM, 2012.

Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proceedings of the international conference on Web search and web data mining*, pages 219–230, 2008.

Thorsten Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge, MA, USA, 1999. URL http://www.bibsonomy.org/bibtex/2596890ca2728ddb94dc9695b19083b82/utahell.

Hyun Duk Kim and ChengXiang Zhai. Generating comparative summaries of contradictory opinions in text. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 385–394. ACM, 2009.

Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods* *in Natural Language Processing*, pages 423–430. Association for Computational Linguistics, 2006.

Kevin Knight and Daniel Marcu. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107, 2002.

Kevin Lerman and Ryan McDonald. Contrastive summarization: an experiment with consumer reviews. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics, companion volume: Short papers*, pages 113–116. Association for Computational Linguistics, 2009.

Kevin Lerman, Sasha Blair-Goldensohn, and Ryan McDonald. Sentiment summarization: Evaluating and learning user preferences. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 514–522, 2009.

Jure Leskovec13, Natasa Milic-Frayling, and Marko Grobelnik. Impact of linguistic analysis on the semantic graph coverage and learning of document extracts. 2005.

Jiwei Li and Sujian Li. A novel feature-based bayesian model for query focused multi-document summarization. *arXiv preprint arXiv:1212.2006*, 2012.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, 2004.

Chin-Yew Lin and Eduard Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics*, volume 1 of *COLING '00*, pages 495–501, 2000.

Chin-Yew Lin and Eduard Hovy. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization-Volume 4*, pages 45–51. Association for Computational Linguistics, 2002.

Chin-Yew Lin, Guihong Cao, Jianfeng Gao, and Jian-Yun Nie. An information-theoretic approach to automatic evaluation of summaries. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 463–470. Association for Computational Linguistics, 2006.

Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM, 2005.

Jingjing Liu, Yunbo Cao, Chin-Yew Lin, Yalou Huang, and Ming Zhou. Low-quality product review detection in opinion summarization. In *The 2007 Conference on Empirical Methods on Natural Language Processing and Computational Natural Language Learning*, pages 334–342, 2007.

Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. Modeling and predicting the helpfulness of online reviews. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 443–452. IEEE, 2008.

Annie Louis and Ani Nenkova. What makes writing great? first experiments on article quality prediction in the science journalism domain. *Transactions of Association for Computational Linguistics*, 2013.

Bin Lu, Myle Ott, Claire Cardie, and Benjamin K Tsou. Multi-aspect sentiment analysis with topic models. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 81–88. IEEE, 2011.

Yue Lu and Chengxiang Zhai. Opinion integration through semi-supervised topic modeling. In *Proceedings of the 17th international conference on World Wide Web*, pages 121–130, 2008.

Yue Lu, Panayiotis Tsaparas, Alexandros Ntoulas, and Livia Polanyi. Exploiting social context for review quality prediction. In *Proceedings of the 19th international conference on World wide web*, pages 691–700, 2010.

Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

Ryan McDonald, Koby Crammer, and Fernando Pereira. Online large-margin training of dependency parsers. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL'05, pages 91–98, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/1219840.1219852.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, pages 171–180, 2007.

Susan M Mudambi and David Schuff. What makes a helpful online review? a study of customer reviews on amazon. com. *Management Information Systems Quarterly*, 34(1): 11, 2010.

Arjun Mukherjee and Bing Liu. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers*, volume 1, 2012.

Courtney Napoles, Benjamin Van Durme, and Chris Callison-Burch. Evaluating sentence compression: Pitfalls and suggested remedies. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 91–97. Association for Computational Linguistics, 2011.

Melissa M. Nelson and Christian D. Schunn. The nature of feedback: how different types of peer feedback affect writing performance. In *Instructional Science*, volume 37, pages 375–401, 2009.

Ani Nenkova and Lucy Vanderwende. The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101*, 2005.

Michael P O'Mahony and Barry Smyth. Using readability tests to predict helpful product reviews. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, pages 164–167, 2010.

Jahna Otterbacher. Inferring gender of movie reviewers: exploiting writing style, content and metadata. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 369–378, 2010.

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, page 71, 2001.

Ana-Maria Popescu, Bao Nguyen, and Oren Etzioni. Opine: Extracting product features and opinions from reviews. In *Proceedings of HLT/EMNLP on interactive demonstrations*, pages 32–33. Association for Computational Linguistics, 2005.

Dragomir R Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6): 919–938, 2004.

Lakshmi Ramachandran, Balaraman Ravindran, and Edward F Gehringer. Determining review coverage by extracting topic sentences using a graph-based clustering approach. 2013.

Martin Riedl and Chris Biemann. How text segmentation algorithms gain from topic models. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 553–557. Association for Computational Linguistics, 2012.

Agnes Sandor and Angela Vorndran. Detecting key sentences for automatic assistance in peer-reviewing research articles in educational sciences. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*, pages 36–44, 2009.

Christina Sauper and Regina Barzilay. In *Automatic Aggregation by Joint Modeling of Aspects and Values*, volume 46, pages 89–127, 2013.

Yohei Seki, Koji Eguchi, Noriko Kando, and Masaki Aono. Opinion-focused summarization and its analysis at duc 2006. In *Proceedings of the Document Understanding Conference (DUC)*, pages 122–130, 2006.

Rachel A Simmons, Peter C Gordon, and Dianne L Chambless. Pronouns in marital interaction what do "you" and "i" say about marital health? *Psychological science*, 16 (12):932–936, 2005.

Anthony Stark, Izhak Shafran, and Jeffrey Kaye. Hello, who is calling?: can words reveal the social nature of conversations? In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 112–119, 2012.

Jiliang Tang, Huiji Gao, Xia Hu, and Huan Liu. Context-aware review helpfulness rating prediction. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 1–8. ACM, 2013.

Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1): 24–54, 2010.

Ivan Titov and Ryan McDonald. A joint model of text and aspect ratings for sentiment summarization. *Urbana*, 51:61801, 2008a.

Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, pages 111–120, 2008b.

Oren Tsur and Ari Rappoport. Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. In *Proceedings of the Third International AAAI Conference on Weblogs and Social Media (ICWSM2009)*, pages 36–44, 2009. URL http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/180.

Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424, 2002.

Aldert Vrij, Samantha Mann, Susanne Kristen, and Ronald P Fisher. Cues to deception and ability to detect lies as a function of police interview styles. *Law and human behavior*, 31(5):499, 2007.

Hanna M Wallach, David M Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In *NIPS*, volume 22, pages 1973–1981, 2009.

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

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354, 2005.

Wenting Xiong and Diane Litman. Identifying problem localization in peer-review feedback. In *Intelligent Tutoring Systems*, pages 429–431. Springer, 2010.

Wenting Xiong and Diane Litman. Automatically predicting peer-review helpfulness. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 502–507, 2011a.

Wenting Xiong and Diane Litman. Understanding differences in perceived peer-review helpfulness using natural language processing. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 10–19. Association for Computational Linguistics, 2011b.

Wenting Xiong and Diane Litman. Evaluating topic-word review analysis for understanding student peer review performance. 2013.

Wenting Xiong and Diane Litman. Empirical analysis of exploiting review helpfulness for extractive summarization of online reviews. In *Proceedings of the 25th conference on Computational linguistics*, 2014.

Wenting Xiong, Diane J Litman, and Christian D Schunn. Assessing reviewer's performance based on mining problem localization in peer-review data. In *EDM*, pages 211–220. ERIC, 2010.

Xiaohui Yu, Yang Liu, Xiangji Huang, and Aijun An. Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE Transactions onKnowledge and Data Engineering*, 24(4):720–734, 2012.

Yi-Ching Zeng and Shih-Hung Wu. Modeling the helpful opinion mining of online consumer reviews as a classification problem. In *IJCNLP 2013 Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 29–35, 2013.

Zhu Zhang. Weighing stars: Aggregating online product reviews for intelligent e-commerce applications. *Intelligent Systems, IEEE*, 23(5):42–49, 2008.

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 43–50. ACM, 2006.