Computer and Information Science; Vol. 11, No. 2; 2018
ISSN 1913-8989 E-ISSN 1913-8997
Published by Canadian Center of Science and Education

Corpus Analysis and Annotation for Helpful Sentences in Product Reviews

Hana Almagrabi[1], Areej Malibari[1] & John McNaught[2]

[1] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

[2] School of Computer Science and National Centre for Text Mining, University of Manchester, Manchester, UK
Correspondence: Areej Malibari, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. E-mail: aamalibari@kau.edu.sa, halmaghrabi@kau.edu.sa, john.mcnaught@manchester.ac.uk

Received: March 8, 2018 Accepted: March 22, 2018 Online Published: April 30, 2018

doi:10.5539/cis.v11n2p76 URL: https://dx.doi.org/10.5539/cis.v11n2p76

Abstract

For the last two decades, various studies on determining the quality of online product reviews have been concerned with the classification of complete documents into helpful or unhelpful classes using supervised learning methods. As in any supervised machine-learning task, a manually annotated corpus is required to train a model. Corpora annotated for helpful product reviews are an important resource for the understanding of what makes online product reviews helpful and of how to rank them according to their quality. However, most corpora for helpfulness are annotated on the document level: the full review. Little attention has been paid to carrying out a deeper analysis of helpful comments in reviews. In this article, a new annotation scheme is proposed to identify helpful sentences from each product review in the dataset. The annotation scheme, guidelines and the inter-annotator agreement scores are presented and discussed. A high level of inter-annotator agreement is obtained, indicating that the annotated corpus is suitable to support subsequent research.

Keywords: product reviews, helpfulness, content analysis, corpus annotation, inter-annotator agreement

1. Introduction

Opinions play a significant factor in people's decision-making. Individuals are influenced by others' advice and evaluations in the process of taking a decision. Word-of-mouth communication is a well-known means to shape consumers' attitudes towards a product (Brown & Reingen, 1987). According to Harrison-Walker (2001), the phrase word-of-mouth refers to: "informal, person-to-person communication between a perceived noncommercial communicator and a receiver regarding a brand, a product, an organization, or a service". The Internet makes it possible for individuals to read about experiences of other individuals, through what is called electronic word-of-mouth (eWOM). Since the emergence of Web 2.0, there has been an explosive growth of eWOM, also called user-generated content (UGC), such as forums, product reviews and web blogs.

Companies want to know consumers' opinions about their product. Potential consumers also want to know opinions from the existing consumers of the product before buying it. Product reviews posted on many e-commerce websites such as Amazon.com are an important type of UGC, enabling companies and individuals to read opinions about a product or a service. However, it is very hard if not impossible for the average person to identify relevant sites and extract and summarize opinions from what are typically very large amounts of reviews. Moreover, product manufacturers also find it hard to identify, manage and summarize opinions from the Web. Therefore, there is a need for automated sentiment analysis systems to provide solutions to manage the abundant opinions posted online.

Sentiment analysis (SA), also known as opinion mining, is a growing field in text mining technology, concerned with the analysis of people's opinions, attitudes, evaluations and emotions expressed in free-text fashion towards different objects such as organizations, product attributes, social topics and individuals (B. Liu, 2012). In SA studies, an opinion can take many forms: a paragraph, a word, sentence or a full review document (Chen & Tseng, 2011). Real-life applications benefit from sentiment analysis studies. For example, models have been proposed to predict sales performance (Yang Liu, Huang, An, & Yu, 2007). Furthermore, opinion mining has been applied to legal blogs, such as in analyzing reactions to high-level court decisions and examining reputations of law firms based on client feedback (Conrad & Schilder, 2007). There has been much research in opinion summarization and polarity identification of positive and negative opinions from product reviews, and major breakthroughs and

promising results have occurred.

However, it is important to provide high-quality content for applications such as sentiment classification and opinion summarization to operate on.

Because of the significance of ranking and classifying reviews based on their quality or helpfulness, most e-commerce sites provide a metric for assessing the helpfulness of reviews, using manual customer feedback on each review. For example, in Amazon.com readers are asked to determine if the review was helpful to them by answering a "Yes" or "No" question. Then the aggregated feedback results are displayed right before each review, e.g., "73 out of 89 people found the following review helpful" (see Fig. 1). Although most review sites provide this manual helpfulness feedback, automatic determination of the helpfulness of reviews is needed for two reasons:

1- Relying on the manual helpfulness feedback of users is unreliable because of three types of bias discovered from the extensive analysis of J. Liu, Cao, Lin, Huang, and Zhou (2007). Section 2 demonstrates details on the types of biases.

2- Improving the results of SA systems. The major weakness of past SA's studies is that all the reviews are treated equally in the analysis, including low-quality reviews. Therefore, this will affect the results of sentiment classification and summary generation.

Many approaches have been developed to automatically assess the quality of product reviews in the literature. Previous studies have been concerned with the classification of complete documents into helpful or unhelpful classes. However, little attention has been paid to performing a deep analysis of helpful review sentences. In this work, we tackle the first step in supervised-learning approach for predicting the quality of reviews' sentences by provide a manually labelled dataset on a fine-grained level, which is at the sentence level. The annotation task aimed to identify helpful sentences in each product review of our chosen dataset.

2. Related Work

Classification methods employed for textual information have been used to organize the vast amount of UGC, for example, emails, publications, product reviews and discussion forums. Most of the work in text classification focuses on classifying a complete document, for example, for sentiment classification (Dave, Lawrence, & Pennock, 2003; Gamon, Aue, Corston-Oliver, & Ringger, 2005; Pang, Lee, & Vaithyanathan, 2002), spam detection in emails (Hotho, Nürnberger, & Paaß, 2005) and assessing the quality of product reviews (Chen & Tseng, 2011; Huang, Shen, Feng, Zhang, & Baudin, 2009; Kim, Pantel, Chklovski, & Pennacchiotti, 2006; J. Liu et al., 2007; Yang Liu, Huang, An, & Yu, 2008; Lu, Tsaparas, Ntoulas, & Polanyi, 2010; Weimer & Gurevych, 2007).

The simplest form of sentiment analysis is to classify a whole review document as having a positive, negative or neutral sentiment about a product (Pang et al., 2002; Turney, 2002). However, reviews often have a mix of opinions about different features of a product. A more fine-grained level of analysis is to determine the sentiment orientation (SO) of each sentence in the document. Classifying sentences into predefined classes has been proposed in sentiment classification research (Hatzivassiloglou & Wiebe, 2000; Kim & Hovy, 2004; Riloff, Patwardhan, & Wiebe, 2006; Wiebe, Wilson, Bruce, Bell, & Martin, 2004; Wilson, Wiebe, & Hwa, 2006; Yu & Hatzivassiloglou, 2003). Some studies have examined the semantic orientation to a topic or product (Hu & Liu, 2004; Popescu & Etzioni, 2007). More information about SA research can be found in B. Liu (2012), Pang and Lee (2008) and Vinodhini and Chandrasekaran (2012). Information about assessing the quality of product reviews can be found in Almagrabi, Malibari, and McNaught (2015).

Motivated by the fact that product reviews vary greatly in quality, many approaches have been developed to assess the quality of product reviews in the literature. Past research typically used the helpfulness score of each review (i.e. manual helpfulness feedback) as the ground truth of their work. Zhang and Varadarajan (2006), used a dataset collected from Amazon.com along with the helpfulness ratio of each review to build a support vector machine (SVM) regression model to estimate the quality of reviews. Kim et al. (2006) also used the helpfulness feedback submitted by readers to measure the quality of reviews using an SVM regression model. However, using the number of helpfulness votes to determine the quality of a review can be problematic.

J. Liu et al. (2007) argued that relying on the helpfulness feedback of users is unreliable because of three types of bias discovered from their extensive analysis: 1) Reviews with high helpfulness score are prominently displayed, this may impact the helpfulness score because of the disproportionate influence on users. This type of bias is referred to as "winner circle" bias; 2) In addition, from an in-depth analysis of Amazon's highly-voted reviews, the study discovered that some of the reviews are not as good quality as the helpfulness voting score indicates. Readers tend to value others' reviews positively, which makes the distribution of helpfulness evaluation skewed towards the helpful vote, giving an "imbalance vote bias"; 3) The last type of bias from Liu's et al. research is called "early

bird bias". This is where the helpfulness voting score may take a long time to accumulate, particularly in newly posted reviews for low-traffic products. Earlier posted reviews are displayed to readers for a longer time than newly posted reviews. Due to such biases, J. Liu et al. (2007) did not use user-helpfulness feedback as the ground-truth in training and testing their model. They used a classification approach to discard noisy and low quality reviews in order to improve opinion summarization. In their work, a manually coded dataset was essential to analyze the utility of reviews without any encountering of the previously mentioned biases. Another limitation of manual helpfulness feedback is the difficulty of spotting fake or shill reviews: such reviews can receive more helpful votes if they are well crafted. The technology blog Gizmodo reported in 2009 that a communication company was paying individuals willing to post positive reviews on Amazon.com to promote its products[1]. Furthermore, spammers can fake helpfulness votes as an advertising strategy (Lau, Zhang, Xia, & Song, 2010).

In any supervised learning method, a labelled dataset is needed to train the classifier. In J. Liu et al. (2007), a set of specifications was proposed for judging the quality of reviews manually. Two annotators were asked to use the specifications as their annotation guidelines to label 40909 reviews on digital cameras crawled from Amazon.com. In the specifications, four categories of reviews were outlined, the "best", the "good", the "fair" and the "bad" reviews. An SVM was used to perform binary classification: the "bad review" category was the low quality class and the remaining categories constituted the high-quality class. After the classification step, only high-quality reviews were used in generating opinion summarization. Table 1 gives the confusion matrix between the annotators. Inter-annotator agreement (IAA) was calculated using Cohen's kappa statistic (Carletta, 1996). The two annotators achieved a high kappa score of 0.8142.

Table 1. Confusion matrix between the annotators (J. Liu et al., 2007)

Annotator 1	Annotator 2				
	Best	Good	Fair	Bad	Total
Best	294	44	2	0	340
Good	66	639	113	0	818
Fair	0	200	1472	113	1785
Bad	1	2	78	1885	1966
Total	361	885	1665	1998	4909

Although Pang and Lee (2008) criticized certain aspects of Liu et al.'s methodology, they were in broad agreement that user-provided utility evaluations of reviews are unreliable.

In later research, Ying Liu, Jin, Ji, Harding, and Fung (2013) introduced an approach to evaluate the utility of online reviews based on the domain user's perspective, such as from the point of view of manufacturing engineers and product designers. Six final year undergraduates in product engineering were asked to label 1000 reviews of mobile phones from Amazon.com. A 5-degree helpfulness evaluation was conducted using categories of "−2", "−1", "0", "1" and "2", where "−2" is the "least helpful" and "2" is the "most helpful". In the experiment, each student had to read and label all the reviews with no provided annotation guidelines. The annotators had to choose the most appropriate helpfulness label according to their own perspective based on their knowledge, training and exposure in design engineering. No statistics are given for IAA, however the authors noted that only just over 1% of reviews were labelled the same by all annotators, there being also large standard deviations for many reviews.

Generally, in text classification studies, researchers have been working on the document, sentence and phrase level. In terms of sentence classification, studies have classified sentences into predefined classes, e.g., subjectivity classification for product reviews. Ghose and Ipeirotis (2007) examined the connection between the helpfulness of a review and its subjectivity. The subjectivity of each sentence was determined using a classifier, and then the standard deviation of the subjectivity score of the sentences in a given review was computed and the results compared with the manually annotated reviews. Two annotators were asked to manually classify each review into categories based on how the reviews influenced their making a purchase decision. The annotators had to answer two broad questions:

1. Is the review informative or not? [answered with "yes" or "no"]

2. If you were interested in buying the product, would the review influence your decision? [1. Yes, positively; 2. Yes, negatively; 3. No; and 4. Uncertain]

[1] http://gizmodo.com/5134652/belkin-employee-sheds-light-on-belkins-supposedly-dirty-practices

The IAA was calculated using the kappa statistic. The results demonstrate agreement with kappa 0.739.

Related work at the sentence level includes, for example, sentence classification by Khoo, Marom, and Albrecht (2006) to classify helpdesk sentences. Their corpus consists of 160 emails between customers and helpdesk operators at Hewlett-Packard. The response emails from the helpdesk contain 1486 sentences. The sentences were labelled according to the Dialog Act Markup in Several Layers (DAMSL) annotation scheme (Core & Allen, 1997). Some examples of the sentence classes are "apology", "instruction", "suggestion" and "thanking". The annotation evaluation shows a high IAA of kappa 0.85.

Wicaksono and Myaeng (2013) crawled a dataset from travel Web forums to study the classification of advice comments. The dataset includes 300 threads containing 5199 sentences, which were randomly chosen from each forum. Two annotators were asked to label sentences into advice or non-advice sentences according to detailed guidelines and a definition of advice. The IAA was kappa 0.76. The level of agreement means that there is a sufficient consensus regarding what advice is, from the proposed definition of advice in Wicaksono and Myaeng (2013).

The work of Jindal and Liu (2006) identifies comparative sentences from product reviews using a dataset collected from customer reviews, forum discussions and random news articles. Two annotators were trained to tag each sentence as one of three comparative sentence types. Furthermore, the identification of conditional sentences and the mining of sentiments from them were proposed by Narayanan, Liu, and Choudhary (2009). They manually annotated 1378 sentences from 5 different forums: Cellphone, Automobile, LCD TV, Audio systems and Medicine. The annotation scheme included tagging conditional and consequent clauses, and identified the product features being commented on and their SO. The annotation guidelines considered only sentences that included at least one sentiment word or phrase. The IAA for sentiment annotation was found to be kappa 0.63.

Sentence classification was also employed to distinguish between qualified claims and bald claims in online product reviews (Arora, Joshi, & Rosé, 2009). To distinguish bald claims from qualified ones, detailed guidelines were established, and annotators were trained. However, to our knowledge, no work was done here to identify helpful sentences from product reviews. In general, a qualified claim can be a fact or a statement that is well-defined and attributed to a source (Arora et al., 2009). Bald claims are non-factual comments and can be open to interpretation, so cannot be verified. This study applied its proposed annotation scheme to the product-review dataset released by Hu and Liu (2004). Two annotators labelled each relevant sentence as being a qualified or a bald claim, with IAA being kappa 0.465. On a separate dataset of 365 review sentences, the agreement was evaluated after removing about 14% of borderline cases. A statistical improvement was established, as kappa rose to 0.532.

In the context of product reviews, we conclude it would be useful to provide a manually annotated corpus for individual sentences that express meaningful information related only to the product. Although there has been increasing interest in the quality prediction of product reviews, there is no available annotated corpus for helpfulness prediction. In this work, we aimed to provide an annotated corpus for assessing the helpfulness of product reviews on the sentence level. We annotated and evaluated our chosen datasets according to the proposed annotation scheme. Two annotators processed each sentence in the corpus.

3. Data and Annotation Procedure

The annotation was performed using the Brat[1] web-based annotation tool. Detailed annotation guidelines were developed from an analytical observation of our data and annotators were trained to recognize helpful sentences. The IAA was then calculated using Cohen's kappa (Landis & Koch, 1977).

3.1 The Dataset

We used the product review dataset[2] released by Hu and Liu (2004), which is freely available for research purposes. The dataset consists of a collection of customer reviews for five different product categories: two digital cameras, one cellular phone, one MP3 player and one DVD player. The reviews were collected from Amazon.com and C|net.com. Each review contains a textual content, a title, and some metadata about the review such as the date, time, and rating. Hu and Liu's work focused on identifying sentences expressing positive or negative opinions towards the features of a product. However, in our proposed annotation scheme we classify all statements from each review, not restricting to sentiment comments alone. Detailed guidelines were provided to our annotators to help them identify the predefined helpful class. If a sentence fell under the specification of a helpful comment, the

[1] http://brat.nlplab.org/

[2] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets

annotators were asked to label it as helpful.

A total of 4035 sentences from 307 reviews of five different products were processed by both annotators independently. Table 2 shows details of our corpus.

Table 2. Product Reviews Dataset

Product Type	Number of Reviews	Number of Sentences
Digital Camera1: Canon G3	46	623
Digital Camera2: Nikon Coolpix 4300	33	365
Cellular Phone: Nokia 6610	40	551
MP3:Player: Creative Labs Nomad Jukebox Zen Xtra 40GB	95	1723
DVD Player: Apex AD2600 Progressive-scan DVD player	96	773
Total	310	4035

3.2 Brat Annotation Tool

The Brat rapid annotation tool offers a collaborative Web-based text annotation environment. Brat supports span-of-text annotations, which makes it an excellent choice for our sentence annotation task. Before we could start our annotation, configuration files had to be created in order to define the types of text spans to annotate. After creating the configuration files, we prepared the data collection of review files. The original product review dataset includes five text files, and each contains a collection of reviews for each product type. In Brat, a text file is needed for each document (review) to be annotated. In addition, for each text document, there is a corresponding annotation file (.ann), both sharing the same name. For example, the file canon1.ann contains the annotations for the review file canon1.text. After setting up the data collection files, we trained a team of two annotators to use the Brat tool. The annotators hold MSc degrees in Finance and Health Management, respectively, from UK universities.

3.3 The Purpose of the Annotation

Automatically assessing the helpfulness of reviews has been studied as a means to help both individuals and companies acquire qualitative reviews quickly. Furthermore, in order to obtain the full advantages of opinion mining systems, it is important to be able to identify high-quality/helpful reviews automatically. In this work, our high-level objective is to perform a deep analysis of review helpful sentences through a corpus annotation task. We introduce detailed annotation guidelines and an annotation scheme that identifies properties of useful comments related only to a product and its features.

Product reviews are not created equal: some short reviews may have more than one comment related to the product and other longer reviews may have only one comment related to the product. We argue that comments containing explicit or implicit mentions of product features and revealing emotions, regular and comparative opinions, product-information (q.v.) and advice about the product being reviewed, are helpful to potential customers and product manufacturers. Therefore, this annotation task aims to identify helpful sentences in each product review.

3.4 Annotation Schem and Guildelines

Research related to product reviews has always investigated relevant information about products, and how this information can be mined using text mining methods. For example, some research related to product reviews has focused on extracting emotional expressions from blogs at sentence level using supervised methods (Das & Bandyopadhyay, 2010). Other work investigates sentiment classification tasks by identifying opinions about products at sentence level. This level of analysis determines the sentiment polarity of each sentence in the review of a product (Yu & Hatzivassiloglou, 2003). However, the results of sentiment classification and emotion identification do not help to identify the reasons behind these emotions and opinions. This gap has motivated researchers to identify reasons for opinions, in order to understand why users like or dislike a product, i.e., focusing on what is termed product-information (Pang & Lee, 2008). Other studies have been concerned with identifying advice-revealing sentences from product reviews to provide insights into the minds of consumers (Ramanand, Bhavsar, & Pedanekar, 2010; Wicaksono & Myaeng, 2013). In addition, researchers have tackled the problem of identifying comparison sentences between products and between their shared features, to help potential customers to consider other options about other products (Jindal and Liu 2006). Our work aims to integrate the previously mentioned research objectives and employ them in a new helpfulness assessment annotation framework based on

experience information (EI) provided by reviewers. In our annotation scheme, we refer to helpful sentences as EI and they can be broadly divided into three types:

1. Sentiment (emotion, regular opinions and comparative opinions)
2. Product-information, and
3. Advice.

We argue that sentences containing explicit or implicit mentions of product features and revealing emotions, regular and comparative opinions, product-information and advice reflect EI and are considered helpful to customers and product manufacturers. The following detailed guidelines were given to the annotators in order for them to distinguish between helpful and unhelpful sentences. Only sentences judged helpful were to be annotated (with the tag Experience-Information), any sentences left unannotated would be taken as unhelpful. Furthermore, examples of the different types of helpful sentences were given. All sentences expressing one of the following meanings were considered helpful:

1. Sentiment: Sentences expressing the reviewer's emotions, reactions and personal taste about the product or one of its features. Opinions about other products are not labelled as helpful because they are not expressing information about the product being reviewed. Opinion comments can be broadly divided into three types:

1.1 Emotions: A self-attributed emotion towards the product or one of its features. Certain keywords can be helpful in spotting emotion sentences such as "disgusted", "angry", "sad", "happy", "surprised", "love", "fearful", "frustrated", "disappointed", "like" and "hate". Some examples are in the following list:

- This is my first digital camera, and I am very pleased with it.
- I recently purchase the canon powershot g3 and I am extremely satisfied with the purchase.

1.2 Regular opinions: Opinions about the product or product-features. Adjectives and adverbs are important keywords to help in spotting opinions such as "bad", "good", "awesome", "great", "well", "fine", "awful", "terrible", "beautiful", "easy", and “difficult". For example the sentence: it is a great phone. Furthermore, words that describe the physical characteristics of the product are indicators for recognizing opinion sentences. Sentences containing the words: "drawback", "pros", "cons" or their synonyms are considered regular opinions. Some examples of annotated regular opinions are:

- Although canon’s batteries are proprietary, they last a really long time, recharge fairly quickly in the camera.
- The only drawback is the viewfinder is slightly blocked by the lens.

1.3 Comparative opinions: Comments where explicit comparison between two or more products or their shared entities is expressed. In general, comparison is expressed by using comparative and superlative adjectives such as "better/the best" or "more difficult/less difficult". There are other words and phrases which indicate a comparison such as "outperform", "number one", "unmatched", "exceed", "prefer", "than" and "the same as". Comparative sentence examples are:

- It’s significantly lighter that the g2 and packed with even more features.
- I have owned Motorola, Panasonic and Nokia phones over the last 8 years and generally preferred Nokia.

2. Product-information: A fact about the product or other information related to its features or service or usage, for example, what features are included in the product or what problems the reviewer reports about the product from their experience. Product-information given by customers reflects their experience of the product. Identifying opinions and emotions about the product is significant for a user, though not enough to take a purchase decision. We note in passing that this type of sentence has not been included in previous research on sentiment analysis, because there is no expressed sentiment. In the following list we show some examples of annotated sentences expressing information about products:

- You can have different kind of lens if you want + flashes etc.
- The only feature missing for me is the voice recognition.
- The prints are beautiful! and you get about 120 images on 256mb card at highest quality.

3. Advice: Expressing suggestions for, or a guide to, an action in certain situations relayed in some context can be helpful. In our annotation scheme, we are interested in explicit advice-revealing sentences. There are some cues to trigger advice, for example, the use of the personal pronoun "you" and modal verbs, e.g., "you should", "you must". Mention of the words "suggestion", "suggest", "recommendation", "advice", etc., can be useful in identifying helpful sentences revealing advice. Here are examples of annotated advice sentences:

- As its 4mp, you might need bigger storage to store high quality images and recording movies (you can record 3 minutes of videos).
- Just double check with customer service to ensure the number provided by amazon is for the city/exchange you wanted.

3.5 General Guildlines

Some general instructions were given for the annotation procedure:

1. Annotators were asked to complete the annotation task independently without discussing the annotation with others.

2. Annotators were asked to take the context into account. For example, some sentences use coreference, thus looking at the wider context would be required to determine if the current sentence should be annotated as EI.

3. Some sentences are part of a list. Annotators were asked to implicitly prepend the header of the list to each point of the list, meaning each point was to be annotated as a sentence. For example, the header *This camera has the following cons:* would be implicitly prepended to each point in the list following it, given below, yielding thus four annotated sentences:

1. Low resolution.
2. No backlit.
3. Short battery life
4. Annotators were asked not to annotate as EI sentences expressing information about other products or brands unless these were expressed in a comparative manner with the product being reviewed.
5. Some sentences express more than one helpful type. Annotators were asked to label these sentences as EI as long as they expressed at least one type. For example, the following sentence expresses a product-information and a regular opinion: the canon g3 gives tons of control for photo buffs but still has an auto mode that makes it very easy for the novice to use.
6. Some sentences report an event related to the product; these comments reflect the customer's experience of the product and, therefore, annotators were asked to label them as EI (e.g., It can't play all of the DVDs).
7. Finally, it is worth recalling that annotation is a very subjective task. With this in mind, the annotators were asked to report any comments during the annotation and these comments were subsequently added to or otherwise incorporated in the guidelines. The guidelines presented in this work are the final version that the authors agreed on. The annotators were asked to re-annotate the whole dataset after all the refinements and changes to our original guidelines.

4. Evaluation

The kappa results we obtained for IAA were interpreted based on the classification of Landis and Koch (1977) which is shown in table 3.

Table 3. Interpretation of kappa

Kappa	Agreement
<0	Less than chance agreement
0	Poor
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate Agreement
0.61 – 0.80	Substantial agreement
0.81 – 0.99	Almost perfect agreement

A total of 4035 sentences were coded by two annotators. The kappa statistic was used to calculate the inter-annotator agreement for each document collection of our dataset. Statistic about the annotation of the DVD player, digital camera 1, digital camera 2, cellular phone, and the MP3 player are shown in tables 4, 5, 6, 7 and 8, respectively. The average kappa score for all collections is given in table 9.

Table 4. The kappa results for the DVD player

		Annotator 1		
		Helpful	Unhelpful	Sum
Annotator 2	Helpful	361	34	395
	Unhelpful	34	344	378
	Sum	395	378	773
	Kappa		**0.82**	

Table 5. The kappa results for the digital camera 1

		Annotator 1		
		Helpful	Unhelpful	Sum
Annotator 2	Helpful	321	29	350
	Unhelpful	20	253	273
	Sum	341	282	623
	Kappa		**0.84**	

Table 6. The kappa results for the digital camera 2

		Annotator 1		
		Helpful	Unhelpful	Sum
Annotator 2	Helpful	200	15	215
	Unhelpful	13	137	150
	Sum	213	152	365
	Kappa		**0.84**	

Table 7. The kappa results for the cellular phone

		Annotator 1		
		Helpful	Unhelpful	Sum
Annotator2	Helpful	282	11	293
	Unhelpful	19	239	258
	Sum	301	250	551
	Kappa		**0.89**	

Table 8. The kappa results for the MP3 player

		Annotator 1		
		Helpful	Unhelpful	Sum
Annotator 2	Helpful	901	36	937

	Unhelpful	46	740	786
	Sum	947	776	1723
	Kappa		**0.90**	

Table 9. The average results of kappa for all document collections

	*Apex AD2600 Progressive-scan **DVD player***	*Canon G3 **Digital Camera1***	*Nikon **Digital Camera2***	*Nokia 6610 **Cellular Phone***	*Creative Labs Jukebox Zen Xtra 40GB **MP3***
P(a)	0.91	0.92	0.92	0.95	0.95
P(e)	0.50	0.51	0.51	0.50	0.50
Kappa	0.82	0.84	0.84	0.89	0.90
Average Kappa			**0.86**		

5. Discussion

The results show a very high average IAA of kappa 0.86. We believe that we have achieved this high score because our guidelines were refined many times taking into account comments from annotators. Initially, our annotation scheme had four categories instead of two, namely emotions, opinions, product-information and advice. However, after starting the annotation process we found that some sentences contain more than one category, for example, emotion, advice and opinion occur in the following sentence: *I love this phone, however I recommend buying an extra battery because it has a low battery life*. It was confusing for annotators to choose only one category to label some sentences. Therefore, we decided to change our annotation scheme to include only two categories, the helpful (annotated as Experience-Information) and the unhelpful. The annotation was re-started so that all sentences were then annotated with the new, reduced scheme This solution helped annotators to identify helpful comments quickly as long as they expressed at least one of our helpful sentence types: emotions, regular and comparative opinions, product information and advice. Clearly, we achieved a high kappa result because there are only two categories to choose from. Moreover, the small number of annotators ensures some consistency, however reading so many reviews improved the knowledge of the annotators regarding the products. Accordingly, we believe that our proposed scheme is more suitable to electronic reviews that include well-defined product-features than to other reviews such as movie and book reviews.

6. Conclusion

Previous work in the field of corpus construction for sentiment analysis of product reviews has mainly been concerned with the manual annotation of positive or negative orientation towards a product. The annotation units include full document (i.e., the review text), sentences and phrases. The extensive amount of user-generated reviews on the Internet has raised concerns about their quality and reliability. Moreover, past research has thrown doubt on the value of helpfulness information typically provided with online reviews when it comes to training models.

We have noted that little attention has been paid to performing a deep analysis of helpful review comments. Previous studies indicated that the subjectivity of a review has a strong effect on utility evaluation (Ghose & Ipeirotis, 2007). However, we argue that subjectivity is not enough for the utility prediction of product reviews. For example, advice-revealing sentences on how to use the product or sentences providing product information with no expressed sentiment are valuable for users in terms of helpfulness.

Our focus has thus been on providing a high-quality/helpful annotated corpus to support the automatic prediction of helpful reviews, with annotation being carried out at the fine-grained level of the sentence. This work is part of a wider project on predicting helpfulness of product reviews, which will, among other things, tackle ranking aspects.

Availability: The annotations and annotation guidelines will be made available via META-SHARE (http://www.meta-share.org) under a CC-BY-NC-SA license.

Acknowledgments

This work is funded as part of a wider project on predicting helpfulness of product reviews. The authors acknowledge with thanks King Abdulaziz University for technical and financial support.

References

Almagrabi, H., Malibari, A., & McNaught, J. (2015). A Survey of Quality Prediction of Product Reviews. *International Journal of Advanced Computer Science and Applications(IJACSA) 11*(6). Retrieved from http://dx.doi.org/10.14569/IJACSA.2015.061107#sthash.RDz4J7sD.dpuf

Arora, S., Joshi, M., & Rosé, C. P. (2009). *Identifying Types of Claims in Online Customer Reviews, Proceedings of the North American Chapter of the Association for Computational Linguistics Azmitia.* Paper presented at the TX: University of Texas.

Brown, J. J., & Reingen, P. H. (1987). Social ties and word-of-mouth referral behavior. *Journal of Consumer research, 14*(3), 350-362.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics, 22*(2), 249-254.

Chen, C. C., & Tseng, Y.-D. (2011). Quality evaluation of product reviews using an information quality framework. *Decision Support Systems, 50*(4), 755-768.

Conrad, J. G., & Schilder, F. (2007). *Opinion mining in legal blogs.* Paper presented at the Proceedings of the 11th international conference on Artificial intelligence and law.

Core, M. G., & Allen, J. (1997). *Coding dialogs with the DAMSL annotation scheme.* Paper presented at the AAAI fall symposium on communicative action in humans and machines.

Das, D., & Bandyopadhyay, S. (2010). *Identifying Emotional Expressions, Intensities and Sentence Level Emotion Tags Using a Supervised Framework.* Paper presented at the PACLIC.

Dave, K., Lawrence, S., & Pennock, D. M. (2003). *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews.* Paper presented at the Proceedings of the 12th international conference on World Wide Web.

Gamon, M., Aue, A., Corston-Oliver, S., & Ringger, E. (2005). *Pulse: Mining customer opinions from free text.* Paper presented at the international symposium on intelligent data analysis.

Ghose, A., & Ipeirotis, P. G. (2007). *Designing novel review ranking systems: predicting the usefulness and impact of reviews.* Paper presented at the Proceedings of the ninth international conference on Electronic commerce.

Harrison-Walker, L. J. (2001). The measurement of word-of-mouth communication and an investigation of service quality and customer commitment as potential antecedents. *Journal of service research, 4*(1), 60-75.

Hatzivassiloglou, V., & Wiebe, J. M. (2000). *Effects of adjective orientation and gradability on sentence subjectivity.* Paper presented at the Proceedings of the 18th conference on Computational linguistics-Volume 1.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). *A brief survey of text mining.* Paper presented at the Ldv Forum.

Hu, M., & Liu, B. (2004). *Mining and summarizing customer reviews.* Paper presented at the Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.

Huang, S., Shen, D., Feng, W., Zhang, Y., & Baudin, C. (2009). *Discovering clues for review quality from author's behaviors on e-commerce sites.* Paper presented at the Proceedings of the 11th International Conference on Electronic Commerce.

Jindal, N., & Liu, B. (2006). *Mining comparative sentences and relations.* Paper presented at the AAAI.

Khoo, A., Marom, Y., & Albrecht, D. (2006). *Experiments with sentence classification.* Paper presented at the Proceedings of the 2006 Australasian language technology workshop.

Kim, S.-M., & Hovy, E. (2004). *Determining the sentiment of opinions.* Paper presented at the Proceedings of the 20th international conference on Computational Linguistics.

Kim, S.-M., Pantel, P., Chklovski, T., & Pennacchiotti, M. (2006). *Automatically assessing review helpfulness.* Paper presented at the Proceedings of the 2006 Conference on empirical methods in natural language processing.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/843571

Lau, R. Y., Zhang, W., Xia, Y., & Song, D. (2010). *Multi-facets quality assessment of online opinionated expressions.* Paper presented at the International Conference on Web Information Systems Engineering.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies, 5*(1), 1-167.

Liu, J., Cao, Y., Lin, C.-Y., Huang, Y., & Zhou, M. (2007). *Low-Quality Product Review Detection in Opinion Summarization.* Paper presented at the EMNLP-CoNLL.

Liu, Y., Huang, X., An, A., & Yu, X. (2007). *ARSA: a sentiment-aware model for predicting sales performance using blogs.* Paper presented at the Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.

Liu, Y., Huang, X., An, A., & Yu, X. (2008). *Modeling and predicting the helpfulness of online reviews.* Paper presented at the 2008 Eighth IEEE International Conference on Data Mining.

Liu, Y., Jin, J., Ji, P., Harding, J. A., & Fung, R. Y. (2013). Identifying helpful online reviews: a product designer's perspective. *Computer-Aided Design, 45*(2), 180-194.

Lu, Y., Tsaparas, P., Ntoulas, A., & Polanyi, L. (2010). *Exploiting social context for review quality prediction.* Paper presented at the Proceedings of the 19th international conference on World wide web.

Narayanan, R., Liu, B., & Choudhary, A. (2009). *Sentiment analysis of conditional sentences.* Paper presented at the Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval, 2*(1-2), 1-135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up?: sentiment classification using machine learning techniques.* Paper presented at the Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10.

Popescu, A.-M., & Etzioni, O. (2007). Extracting product features and opinions from reviews *Natural language processing and text mining* (pp. 9-28): Springer.

Ramanand, J., Bhavsar, K., & Pedanekar, N. (2010). *Wishful thinking: finding suggestions and'buy'wishes from product reviews.* Paper presented at the Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.

Riloff, E., Patwardhan, S., & Wiebe, J. (2006). *Feature subsumption for opinion analysis.* Paper presented at the Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing.

Turney, P. D. (2002). *Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews.* Paper presented at the Proceedings of the 40th annual meeting on association for computational linguistics.

Vinodhini, G., & Chandrasekaran, R. (2012). Sentiment analysis and opinion mining: a survey. *International Journal, 2*(6).

Weimer, M., & Gurevych, I. (2007). *Predicting the perceived quality of web forum posts.* Paper presented at the Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP).

Wicaksono, A. F., & Myaeng, S.-H. (2013). *Automatic extraction of advice-revealing sentences foradvice mining from online forums.* Paper presented at the Proceedings of the seventh international conference on Knowledge capture.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning subjective language. *Computational linguistics, 30*(3), 277-308.

Wilson, T., Wiebe, J., & Hwa, R. (2006). Recognizing strong and weak opinion clauses. *Computational Intelligence, 22*(2), 73-99.

Yu, H., & Hatzivassiloglou, V. (2003). *Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences.* Paper presented at the Proceedings of the 2003 conference on Empirical methods in natural language processing.

Zhang, Z., & Varadarajan, B. (2006). *Utility scoring of product reviews.* Paper presented at the Proceedings of the 15th ACM international conference on Information and knowledge management.