A Survey of Quality Prediction of Product Reviews

H.Almagrabi
School of Computer Science
University of Manchester
Manchester, UK

A. Malibari
Dept. of Computer Science
King Abdulaziz University
Jeddah, Saudi Arabia

J. McNaught
National Centre for Text Mining
University of Manchester,
Manchester, UK

***Abstract*—With the help of Web-2.0, the Internet offers a vast amount of reviews on many topics and in different domains. This has led to an explosive growth of product reviews and customer feedback, which presents the problem of how to handle the abundant volume of data. It is an expensive and time-consuming task to analyze this huge content of opinions. Therefore, the need for automated sentiment analysis systems is vital. However, these systems encounter many challenges; assessing the content quality of the posted opinions is an important area of study that is related to sentiment analysis. Currently, review helpfulness is assessed manually; however the task of automatically assessing it has gained more attention in recent years. This paper provides a survey of approaches to the challenge of identifying the content quality of product reviews.**

Keywords—sentiment analysis; product reviews; content analysis; helpfulness detection

I. Introduction

The Internet has made it possible to discover opinions of others on a wide range of subjects, through social media websites, such as review sites and wikis, and through online social networks. According to a survey, 81% of Internet users have done research on a product at least once [1]. Studies have found that customers' reviews can form others' opinions and subsequently affect sales [2, 3, 4].

Understanding and analyzing public opinion is important for the prediction of future events. Consequently, this aids the process of making a decision that can involve improving services, handling political elections and calculating risk management. Organizations conduct consumer surveys to explore opinions about their products and/or services. However, the design and the supervision of these surveys are expensive, tedious and a time-consuming task [5]. It is easier for companies to utilize the freely available online consumer reviews. However, the explosive growth of opinion text on the Web makes it hard to manage. In addition, opinions posted on the Web in free-text style are less structured than those conducted from consumer surveys and focus groups, and require more effort to collect and analyze [6, 7, 8].

Sentiment analysis, also referred to as opinion mining, is a growing field in text mining technology, which is 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, public events and even individuals [9]. Sentiment analysis is motivated by the fact that individuals and organizations are increasingly using the content of social media for decision-making. Since 2000, there has been much attention paid to sentiment analysis research, mainly because of the rise of machine learning (ML) technology in natural language processing (NLP). In addition, the public datasets available for training using ML algorithms have aided sentiment analysis research. There have been major breakthroughs and promising results in research into sentiment analysis, especially in opinion summarization, feature extraction and polarity identification. However, the extensive amount of uncontrolled user-generated reviews on the Internet has raised concerns about their quality and reliability.

There are many challenges in sentiment analysis research such as dealing with sarcasm and implicit opinions, domain dependency, subjectivity detection and entity identification. A significant challenge that has been studied is determining the quality, also called utility, helpfulness or usefulness, of product reviews [10, 11, 12, 13]. This survey will point out some of the significant research to handle the quality prediction problem.

The rest of this paper is organized as follows. In section 2 we have introduced the challenge of 'product reviews quality. Sections 3 and 4 give an overview of the methods used to predict the utility of product reviews. Section 5 discuses some of the research related to the quality prediction problem, while section 6 concludes the paper.

II. Quality of Reviews

The topic of quality of reviews is related to opinion spam detection, which makes it an important area of research. However, according to Liu [9], it is different from spam detection, as spam reviews may not be of low quality. Fake reviews may be of high quality, especially if they are well written, which makes them hard to identify. Determining good quality reviews saves readers time and effort by discarding noisy and low quality content. It is useful to have a mechanism to automatically assess a review's helpfulness as soon as it is written.

Some aggregation and hosting websites rank reviews according to their perceived helpfulness by readers, such as Amazon.com, Epinions, IMDB CitySearch, etc. Users manually assess reviews by responding to a question, such as "Was this review helpful to you?". Readers can respond with "YES or NO" and the feedback results are calculated and displayed next to each review (e.g., "12 of 20 people found the following review helpful"). Although this helpfulness evaluation method has been used by many websites, it is still a meaningful task to automatically determine the quality of each review for the following reasons:

1) Many reviews have little or no helpfulness evaluation, especially in low-traffic items. According to [11], some

reviews lack a helpfulness evaluation: 38% of 20,000 Amazon MP3-player reviews only received three or less votes in three months [11]. In addition, consumers are not obligated to respond to the feedback question to determine a review's perceived helpfulness, even if they found it helpful or not.

2) Human generated helpfulness evaluation may be fake, which makes the helpfulness voting score untrustworthy [1]. Spammers can click on the helpfulness voting buttons (Yes or No) to increase or decrease the helpfulness of a review. Therefore, depending on the helpfulness feedback to identify helpful reviews can be problematic.

3) Biases can be found in the manual helpfulness evaluations [10]. Reviews with high helpfulness score are prominently displayed, which would have a disproportionate influence on readers and consequently on the helpfulness voting score itself. This type of bias is referred to as "winner circle" bias in [10]. In addition, an in-depth analysis of Amazon's highly-voted reviews, lead to discover that some of the reviews are not of 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, known as the "imbalance vote bias". The third type of bias identified is called "early bird bias" [10]. The helpfulness voting score may take a long time to accumulate, particularly in newly posted reviews. Earlier posted reviews are displayed to readers for a longer time than newly posted reviews.

4) The use of robust review quality prediction systems will facilitate ranking reviews according to their utility, and thus users can easily and quickly access them. Furthermore, applications such as sentiment extraction and opinion summarization will benefit from such systems by operating on high quality content rather than spammed and misleading reviews. For example, in the process of opinion summarization it is useful to only use good quality reviews and discard low quality ones, including reviews with high helpfulness voting score, which are subject to the previously mentioned biases. Therefore, automatically classifying reviews according to their quality would aid and speed up the quality of opinion summarization [10].

III. Quality as Regression Problem

Generally, the problem of determining the quality of reviews is seen as a regression problem. The method uses machine learning models to assign a score to each review. These scores can be used in recommendation and ranking systems [9]. Researchers have used different types of features to train and test models on datasets from different domains. A Support Vector Machine (SVM) regression model to rank reviews according to their helpfulness was employed by [11]. They used structural features (e.g., review length, number of sentences), lexical features (e.g., unigrams and bigrams), syntactic features (e.g., nouns, verbs, etc.), semantic features (e.g. sentiment words) and meta-data features (e.g., number of stars) [9]. The most useful features used were the length of the review, its unigram and its product rating.

Zhang and Varadarajan [13] use a similar feature set to that proposed in [11]. However, they did not include any meta-data information. Their study assumes that a good quality review should discuss many aspects of the product. Thus, a comparison between the review and the product specifications was considered. However, the results show that this feature did not improve system performance [1]. Furthermore, the study includes review similarity to editorial reviews, which did not improve the system performance either. They found that the perceived helpfulness in product reviews depends greatly on its linguistic style (e.g., word count, comparatives and superlatives, proper nouns, etc). In contrast with the results of [11], there is a weak correlation between review length and utility score [1, 13]. According to Pang and Lee [1], the difference in domain choice affected the results in the two studies. Electronic product reviews used in [11] do not include as sophisticated language as found in book and movie reviews, which were used in [13].

A different approach was proposed by Ghose and Ipeirotis [14], who studied the relationship between the subjectivity of a review and its helpfulness. A classifier determines the subjectivity of a sentence, and then the standard deviation of the subjectivity score of the sentences in a given review is computed. The results indicate that the standard deviation score and a readability score have a strong effect on utility evaluation. Building on their previous research, Ghose and Ipeirotis [15] expanded their work by examining multiple product categories and by adding textual features, such as history information about the author, readability metrics and spelling errors, etc. They found that reviews including a mixture of subjective and objective information influence sales and the perceived usefulness. In addition, readability and informativeness features were found to correlate positively with sales and the perceived usefulness. An important finding of this research is that the type of product affects the perceived helpfulness of a review. For feature-based products (e.g., electronics), reviews that include objective more than subjective information increase the usefulness of the review. However, in experience products (e.g., movies), it was found that subjectivity matters the most, as users prefer to read personalized and highly sentimental comments that describe the reviewer experience and provide more information about the product.

Looking at the problem from a different perspective than the above approaches, the work in [16] introduced three main factors affecting the helpfulness of a review: reviewer's expertise, review timeline and review style based on part-of-speech tags. A nonlinear regression model was used to integrate the proposed factors. Extensive experiments on movie reviews (IMDB data-set) show the efficiency of the proposed model. They argue that their model is general enough to be employed in other domains, by replacing the genres of movies with the categories of products and by modeling the timelines and the writing style using their proposed algorithm.

Previous research efforts focus only on the meta-data and on the review text itself to analyze various properties of product reviews in order to predict quality. Other studies tried to tackle the quality problem from different perspectives.

For example, the study by [17] incorporates another information resource: the author's behaviour on e-commerce sites, such that information derived from their online transactions helped to assess the quality of reviews and to identify spammed ones. Three features were used to assess the quality of reviews: personal reputation, seller degree and expertise degree. The correlation between each feature and the helpfulness votes was examined using a linear regression analysis.

Lu, Tsaparas, Ntoulas, and Polanyi [18], investigate if the social context of reviews can enhance the performance of quality prediction. In their view, important information can be obtained from the social context about the quality of reviewers, which affects the quality of their reviews. In order to incorporate social context in predicting review quality, regularization constraints, based on a set of experimentally-validated hypotheses, were employed. An example is the "author constraint hypothesis", which assumes that reviews from the same author are similar in quality [18]. The results show accuracy improvement in predicting the quality of reviews using a text-based classifier (linear regression model). They argue that the proposed regularization technique can be applicable and generalized for quality evaluation of other user generated content. However, this method cannot be employed to review sites that do not have a trusted social network [9].

Previous studies have proven that online reviews affect the sales of products, however many studies fail to consider the quality of reviews. Another group of interested researchers [19] proposed a regression model that incorporates the quality factor for predicting sales performance of products being reviewed. Results indicate the positive correlation between review quality and prediction accuracy of sales performance.

IV. Classification and Other Methods

In addition to ranking reviews according to their quality and utility, researchers also used classification methods to determine the quality of reviews. In most previous studies, helpfulness votes were used as the ground-truth data for training and testing regression models [11, 14]. However, in a different approach, these approaches were unreliable because of the previously mentioned types of biases discovered from their extensive analysis [10]. Thus, they did not use user-helpfulness feedback (helpfulness votes) as the ground-truth in training and testing their model. Their work focused on improving the quality of opinion summarization by detecting and discarding noisy and low quality reviews using a classification based approach. The proposed approach explores three features of product reviews: readability, informativeness and subjectivity. A set of specifications was proposed for judging the quality of reviews, and four categories defined: "best reviews", "good reviews", "fair reviews" and "bad reviews". A SVM was used to perform binary classification, with the "bad review" category as the low quality class and the remaining categories as the high quality class. After the classification step, only high quality reviews were used in generating opinion summarization.

O'Mahony and Smyth [20] proposed a classification-based recommender system to recommend the most helpful reviews to the end user. Many features were used to train a classifier to distinguish between helpful and non-helpful reviews: reputation, content, social and sentiment features. The reputation and sentiment features achieved better classification performance than content and social features. A significant finding was that the classification performance remained high, even in the absence of the reputation features, which are not always available.

Chien and Tseng [21] treated the quality problem as a classification problem by employing a multiclass SVM model to classify product reviews. In order to derive informative review features, an information quality "IQ" framework adopted from [22] was employed. Table 1 illustrates the information quality categories along with their dimensions. The authors defined five classes of quality: high, medium, low, duplicated and spam, and the specifications of review quality were adopted from [10]. Furthermore, factors that shape high quality reviews were analyzed and the findings show that helpful reviews need to be subjective and provide detailed comments on a number of product aspects.

TABLE I. Wang and Strong's IQ Framework [22]

IQ Category	IQ Dimensions
Intrinsic IQ	Believability, accuracy, objectivity, reputation
Contextual IQ	Value-added, relevancy, timeliness, completeness, appropriate amount of information
Representational IQ	Interpretability, ease of understanding, representational consistency, concise representation
Accessibility IQ	Accessibility, access security

In a recent study, Bayesian inference was used to measure the probabilities of the reviews belonging to certain classes [72]. In addition, an extended fuzzy associative classifier was developed to train a review helpfulness classification model. The model incorporated features from previous studies [10, 21, 32, 52, 54, 69], for example, subjectivity features, emotion features and stylistic features.

In a different approach from the previous supervised methods, Tsur and Rappoport [23] introduced a fully unsupervised method to rank book reviews according to review helpfulness. First, the proposed REVRANK algorithm identifies the dominant terms in a set of review documents. These important terms represent a "virtual optimal" or a core review representation. Subsequently, reviews are mapped to this optimal representation and a ranking score is given to each review according to distance between the review and the virtual review. All reviews of a given book were explored to generate a lexicon of the dominant concepts.

This is relevant to keyphrase extraction proposed in the TextRank and the CollabRank systems [24, 25]. Both systems employ a graph-based unsupervised ranking algorithm which ranks keyphrases, using the co-occurrence links between words in the TextRank system and the collaborative knowledge from multiple documents in the CollabRank system.

In recent work, a new problem of personalized review quality prediction was addressed to recommend helpful reviews [26]. The authors argue that the quality of reviews may

not be the same for different readers, while all the previous studies assume that it is. They found that there are some latent features that affect the user's evaluation of the quality of the review. Based on this assumption, a series of probabilistic graphical models, based on matrix factorization and tensor factorization, were proposed. The experiment was conducted on a real-life dataset from Eopinion.com, and the results show that the proposed technique outperformed the existing state-of-the-art approaches, at that time, using textual and social features.

All the above studies did not consider that highly ranked reviews may include highly redundant information [9]. Another method was proposed to solve this problem by selecting a small comprehensive set of high quality reviews [27]. These reviews cover many different aspects and viewpoints of the reviewed product. The authors of [27] extended existing algorithms for maximizing coverage to handle this problem. Their work is different in that they selected a set of comprehensive reviews rather than scoring each review. Furthermore, the proposed approach is different from opinion summarization because it aims to identify a subset of reviews that cover the different aspects of a product rather than summarizing the opinions on the extracted features of a product. The most related work to [27] is the work of Lappas and Gunopulos [28]. However, whereas the goal of the former authors is to cover the product aspect from a fixed size set with both negative and positive opinions, that of the latter is to cover all product aspects while preserving opinion distribution.

Miao, Li, and Dai [29], introduced a sentiment mining and retrieval system which is concerned with mining useful information from customers' reviews. They employed both data mining and information retrieval techniques to build a novel temporal opinion quality and relevance ranking system, which mines customers' preferences.

Wu, Greene, and Cunningham [30] compared two aggregation methods for combining sets of features in order to identify untruthful opinions about hotels. Their solution was to build a useful suspicious-review ranking system. The results show that the best features to identify suspicious reviews are: proportion of positive singleton reviews, truncated rating, and reactive positive singleton reviews. Furthermore, it was found that singular value decomposition outperforms the unsupervised hedge algorithm for combining features to identify suspicious reviews about hotels. Although this work falls under spam detection, it is related to identifying qualitative reviews. Determining criteria for identifying suspicious reviews would improve the identification of reliable and trustworthy reviews.

Lau, Zhang, Xia, and Song [31] proposed a method to detect non-informative online opinionated expressions. The proposed multi-facet quality metric utilizes both the intrinsic properties of opinionated expressions and association with other opinionated expressions posted on the Internet. Furthermore, to avoid the biases mentioned in [10], helpfulness votes were not used as the ground-truth for quality assessment.

A novel approach to assess the quality of product reviews was proposed by Min and Park [32]. The proposed metric employs linguistic clues to capture time expressions related to the use of the product and product aspects during different purchase times. They found that tense and time expressions are the most useful linguistic clues to assess the customer's previous purchase experiences. This approach is similar to the work done by the group in [18], however there is a difference, because this work uses the reviewer's social context information based on the social network-based review website, such as the PageRank score of the author. Features used in this work improve system performance, however, they have limitations similar to the “helpfulness votes”. In contrast, the method proposed by Min and Park extracts the reviewer's characteristics directly from the textual content of a review by utilizing his/her experience.

The authors in [33] proposed a method to evaluate the helpfulness of online reviews based on the domain user's perspective, such as manufacturing engineers and product designers. They conducted an exploratory study to understand what makes reviews useful to designers. Four categories of features were proposed, based on their experiment, to identify helpful reviews: product features, linguistic features, features using information theory and features based on information quality. Machine learning algorithms were employed using both classification and regression to evaluate the proposed method. The results show a strong correlation between the designer’s rating and the proposed method.

Another classification approach to modeling the helpfulness problem was proposed by Zeng and Wu [34]. A three-class classification framework was introduced to find the helpful positive reviews, the helpful negative reviews and to filter out the unhelpful reviews. Table 2 lists the features used in the classification approach. Some of the features were adopted from [11] and other features were added based on the findings of [35]. The study uses the list of common ideas related to helpfulness and unhelpfulness proposed in [35]. The performance of the three-class problem is quite high and the results show that helpful reviews (positive and negative) can be identified with high precision from unhelpful ones [35].

Recently, the work of Krishnamoorthy [69] has developed a new method for extracting linguistic features based on the linguistic category model (LCM), proposed by Semin and Fiedler [70]. A binary classifier was built and evaluated using the LCM with a combination of other features, namely: metadata features (e.g., review extremity), readability features, and subjectivity features (i.e., the total number of subjective words normalized by review length). The experiment on two real-life review datasets shows that linguistic category features are better predictors for the helpfulness of product reviews. Table 3 presents the linguistic category features used in the LCM.

TABLE II. CLASSIFICATION FEATURES OF ZENG AND WU [34]

Features	Description
Unigram (Product Description)	The number of unigrams used between the review and the corresponding product description
Bigram (Product Description)	The number of bigram used between the review and the corresponding product description
Trigram (Product Description)	The number of trigrams used between the review and the corresponding product description
Length	The length of a review
Comparisons	The review uses the string "compare to" or "ADJ + er than"
Degree of detail	Defined by formula
Use of Ratings	The "Star" ratings of the review
Pros and Cons	The review contains exactly the strings "Pros" and "Cons"

TABLE III. LINGUISTIC CATEGORIES AND THEIR DESCRIPTION [69]

Category	Description
Adjectives (ADJ)	Qualifies a noun; highly subjective and abstract
SV (State verbs)	Refers to mental or emotional state
SAV (State Action Verbs)	Describes the emotional consequences of an action; high positive or negative connotation
IAV (Interpretive Action Verbs	Multitude of actions that have the same meaning; have a positive or negative connotation
DAV (Descriptive Action Verbs)	Objective description of a specific action; no positive/negative connotation

V. RELATED TASKS

Researchers from many fields have shown great enthusiasm in studying and analyzing the quality of online reviews [10, 11, 13, 14, 16, 18, 36]. Pang and Lee [37] carried out the first study related to this problem. They studied the prediction of product rating, which correlates with the perceived helpfulness of reviews [11]. Automatically scoring essays is another related study which has been used to rate the quality of reviews [38]. The work group of [11] built a regression model to rank reviews according to their quality, employing the same set of features used in essay scoring by [38].

By using statistical methods, Jindal and Liu [39] discovered that the extracted features from reviews, such as rating and title length, improved the identification of spammed and duplicated reviews. The task of evaluating the quality of Web posts and the quality of answers in question and answering systems is also related to predicting the helpfulness of reviews [40, 41, 42, 43, 44, 45]. A study by Hoe, Li, and Zou [46] was to determine whether a review would achieve helpfulness votes, or not, through examining the posting time and textual features of a review. The work in [44] integrates user and community features with the review textual features to assess the quality of questions and answers using classification methods. The model utilizes community features which were driven from the answer text, for example the length of the answer and number of points received, in addition to the user's features, such as number of answers given. The authors of [45] introduce a co-training method to model the quality of both the review and the reviewer in an attempt to use community information for extracting features.

The identification of high quality reviews influences feedback and reputation systems. It was found that seller and buyer behaviour changes in response to the change in a seller's feedback profile [47]. In another similar study [48], it was found that personal reputation positively influences buyers to purchase a product, thus a small amount of negative feedback will not affect sales. However, most research on the quality of reviews analyzes the personal reputation impact on the review itself, rather than on the seller's/buyer's transactions [17]. In addition, the work in [49] has also examined the impact of high quality reviews on purchase decisions and thus on sales. The results show that reviews with a high proportion of helpful votes are more important in making purchase decisions than the aggregate star rating of a product. Another important finding is that the reviewer's reputation has no impact on the consumer purchase decision, which contradicts the findings of [17].

The work in [19] proposed a regression model for predicting sales performance using product reviews. This was one of the first attempts to examine the economic impacts of review quality. The authors of [50] studied the problem of product review search. The retrieved reviews are ranked according to their quality based on a given query. Other studies tackle the problem of finding good answers from question-answering systems such as Yahoo! Answers[1]. A method of determining the quality of the answer was proposed by [51], which considered both textual features from the answers and social features from the answerer.

Some studies are interested in determining the fundamental characteristics of helpful reviews from a theoretical and practical point of view [52, 53, 54, 55, 71]. For example, Mudambi and Schuff examined factors contributing to review helpfulness, by employing a linear regression model. The study shows that product type, review extremity and review depth affect review helpfulness [52]. They analyzed the different impact of product type (search or experience goods) on review helpfulness in the multistage consumer decision process. Search goods are defined as products about which consumers can easily obtain information about their quality before purchasing. Experience goods are products that require purchasing in order to evaluate their quality, for example, books and movies. Thus, the type of product affects a consumer's textual evaluation of a product and consequently affects the perceived helpfulness of that review. The study in [54] examines what factors affect the number of helpfulness votes reviews receive. It is argued that knowing the most important factors to attract helpfulness votes will help website designers to improve their helpfulness voting mechanism. For example, the findings show that semantic features have the most impact on the number of helpfulness votes. The conclusion is that websites should provide more ranking options to rank reviews rather than ranking the most recent review. Siering and Muntermann [55] confirm that product type influences the perceived helpfulness as was discovered in [52]. Furthermore, they find that perceived helpfulness is affected by other textual aspects related to review sentiment, product quality and review uncertainty. Lee and Choeh [71]

[1] http://answer.yahoo.com/

employed a multilayer perceptron neural network model to improve helpfulness prediction accuracy. The product review metadata and the textual characteristics of both the review and the reviewer were used as features in the proposed model. The study found that the list price and the sales rank of the product are important for helpfulness prediction. In addition, the results show that the neural network model outperforms the conventional regression models in helpfulness prediction.

Other studies have discovered other factors that impact review helpfulness, such as specific emotions [56, 57] and review readability [15, 53]. From a theoretical perspective, the authors of [53] proposed a model to investigate the relationship between the textual content of a review and its perceived helpfulness votes. Four readability measures were employed to validate the proposed model. Figure 1 demonstrates the four readability measures. The model was based on three specific aspects: conformity, understandability and extensiveness. The results show that helpfulness is affected by review readability more than its length and that reviews with extreme votes receive a higher score than the less helpful ones. Some studies assume that focusing on product quality in online reviews will provide diagnostic value and thus impact the review helpfulness [58, 59].

Racherla and Friske [60] explore which factors contribute the most to consumer perception of the utility of online reviews. Furthermore, they investigate if the impact of these factors varies according to the type of service offered to the consumer. An important finding of this research is that reviews provided by reviewers with high expertise and good reputation are significantly helpful reviews. In addition, reviews with a great amount of information are not particularly considered more helpful than reviews containing less information.

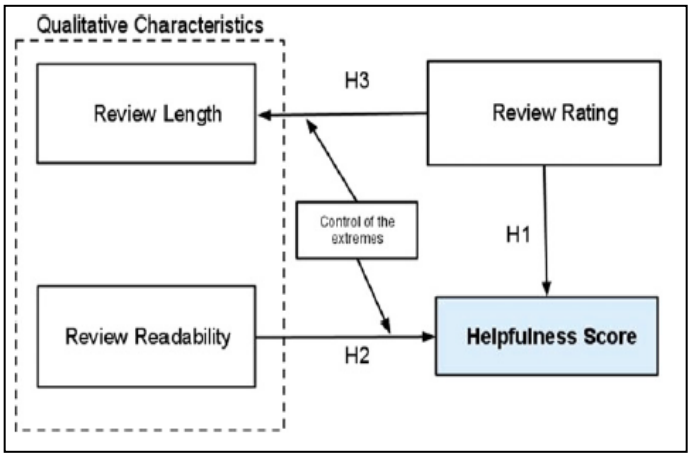

Fig. 1. Theoretical model [53]

The work of Otterbacher [61] has focused on understanding helpfulness rather than just predicting it. A well-established framework for quality assessment by Wang and Strong [22] is adopted. The framework includes four categories of data quality developed from the end user's perspective: Intrinsic quality, Contextual quality, Representational quality and Accessibility. Each category contains several dimensions. Only the first three categories are used because accessibility is not an issue as reviews are accessible in the virtual community environment [22]. A simple linear regression analysis shows that the helpfulness score correlates with several data quality dimensions.

In another study to identify the specification of a helpful review, Connors, Mudambi, and Schuff [35] conducted an open-ended analysis. A group of 40 participants were asked to rate reviews and describe what factors they found in helpful and unhelpful reviews. Table 4 shows the results of the analysis. For example, the "product usage information" was used 30 times by all participants. Categorizing these textual descriptions highlighted the influence of three criteria for helpfulness: the credibility of the reviewer, the similarity between the author and the reader, and the use of positive and negative comments about the product.

In the next stage of their study, a controlled experiment was conducted to study the effect of the three criteria on the helpfulness rating. The findings show that both the characteristics of the reviewer and of the review influence the reader's understanding of review helpfulness. For example, the study recommends that reviewers should disclose their self-described expertise as part of their review. Furthermore, review sites must favour content over balance, a review expressing both positive and negative statements towards a product is not perceived as more helpful as a one-sided review. This idea was also proposed by Schlosser [62], who argues that reviews including two-sided arguments (pros and cons), are not necessarily more helpful, credible and persuasive than a one-sided argument review. The study proves that reviews with two-(versus one-) sided arguments receive higher helpfulness votes only if the reviewer rating is fairly favourable, while reviews written by extremely favourable reviewers are considered more helpful even if they include only one-sided argument.

While preceding studies have examined review characteristics, Ngo-Ye and Sinha [73] incorporate review related engagement features in their proposed hybrid text-regression models to predict review helpfulness. Furthermore, this study uses a bag-of-words representation as part of the textual features. The proposed hybrid model including the textual features and the reviewer engagement characteristics enhanced helpfulness prediction. This work offers new factors that contribute to the prediction of helpfulness. In another attempt to examine the factors affecting the helpfulness of reviews, a recent study by Liu and Park [74] suggested that the combination of the messenger and message features correlates positively with the helpfulness prediction of reviews.

Features such as the reviewer identity and the length of the review were used to build a textual regression model to predict helpfulness. Valence (positive or negative) consistency, a new aspect to the problem of utility prediction, was recently investigated in [75]. That study examined the influence of other nearby reviews on the perceived helpfulness of the review itself. The results show that, whether or not the reviews are being positive or negative, consistent reviews are more helpful than inconsistent ones, which was not the case in prior studies [76, 77]. For example, Scholz and Dorner [78] found that positive reviews achieve better helpfulness scores than negative ones.

TABLE IV. IDEAS RELATED TO HELPFULNESS AND UNHELPFULNESS [35]

Helpfulness	Times Mentioned
Pros and Cons	36
Product Usage	30
Information Detail	24
Good Writing Style	13
Background Knowledge of product	12
Personal Information About Reviewer	12
Comparisons	10
Lay-Man's Terms	9
Conciseness	8
Lengthy	7
Use of Ratings	7
Authenticity	5
Honesty	5
Miscellaneous	4
Unbiased	4
Accuracy	3
Relevancy	3
Thoroughness	3
Unhelpfulness	**Times Mentioned**
Overly Emotional/Biased	24
Lack of Information	17
Irrelevant Comments	9
Not Enough Detail	6
Poor Writing Style	6
Using Technical Language	6
Low Credibility	5
Problems With Quantitative Rating	5
Too Much Detail	5
Too Short	4

Martin and Pu [79] suggested that emotional words are powerful parameters in helpfulness prediction. They propose a framework to extract the emotionality from the textual content of reviews. GALC, a general lexicon of emotional words, was employed to represent a model of 20 categories using supervised classification methods. The results show that the emotion-based method outperforms the previous structural-based methods by 9%. The work of Mertz, Korfiatis, and Zicari [80] examined the helpfulness prediction problem by evaluating the performance of dependency bigrams and discourse connectives. A binary classifier was introduced using the previously mentioned novel text-based features. This study shows that various types of discourse relations are useful set features for predicting review helpfulness. Moreover, there is a strong correlation between high star ratings and helpful reviews. Another study has investigated how misalignment between the star rating and the textual content of the review can lower the overall helpfulness of the review [81]. It found that misalignment between star rating and review text often occurs in reviews of experience goods and in reviews with high star ratings. This theoretical analysis suggests that highly rated experience goods reviews are perceived more helpful than other reviews.

Recently, a study of factors contributing to online review helpfulness was carried out [82]. Specifically, the goal of the study is to examine the joint effect of the message length of a review (word count) together with reviewer characteristics and the patterns of the review on the utility of the review. While prior studies have suggested that there is a positive correlation between word count and the perceived helpfulness of a review [52, 77], the results of this study point out that the association between word count and helpfulness is valid only in reviews with 144 or less words [82]. The hypothesis results of this study are listed in table 5.

TABLE V. RESULTS OF HYPOTHESIS TESTING [82]

H1a: For reviews written by all reviewers, word count is a significant predictor of review helpfulness when the review is shorter than average	Supported
H1b: For the reviews written by top reviewers, word count is a significant predictor of review helpfulness	Not supported
H2: For top reviewers, reviewer experience is a significant predictor of review helpfulness	Not supported
H3: For top reviewers, reviewer impact is a significant predictor of review helpfulness	Not supported
H4: For top reviewers, reviewer cumulative helpfulness is a significant predictor of review helpfulness	Supported
H5: For the top reviewers, product rating is a significant predictor of review helpfulness	Supported

In an approach proposed recently by Tang, Gao, Hu, and Liu [63], the prediction of review helpfulness for each user was investigated. A context-aware helpfulness prediction framework utilizes both the social and the content context of a review. The review content affects the perceived helpfulness by other users. Furthermore, information about the author, rater and their relationship can provide the social context of reviews. For example, it is more likely that raters find reviews from their connected authors helpful. However, this framework has some limitations that need to be addressed. The social context is likely to change over time, such as user preference.

VI. CONCLUSION

With e-commerce websites growing rapidly, reviews about products are becoming an important source of information for making informed purchase decisions. Sentiment analysis research is concerned with extracting and summarizing the massive content of product reviews. However, the explosive growth of product reviews raises concerns about their reliability and quality. Although review helpfulness is currently assessed manually by many retailer websites, it is important to automatically assess the quality of reviews at least for two reasons. The first reason is to provide helpfulness evaluation when human evaluations are lacking. The second reason is to correct skews in human helpfulness evaluations mentioned in [10].

Regression and classification methods have been examined to rank and classify reviews according to their helpfulness. Major challenges in quality prediction include feature weighting, which affects the performance of classifiers enormously.

Another challenge is to consider the quality of the reviewer. Some trust metrics from other studies may be of use to determine the helpfulness of product reviews, for example, research into peer-to-peer and reputation networks [64, 65, 66, 67, 68].

Investigating what factors determine review helpfulness could improve review systems. Therefore, retailers' websites could introduce automatic helpfulness scoring systems to reduce customers' search cost. This would affect customers' satisfaction and purchase behaviour [49, 78]. Furthermore, there should be a method to increase the efficiency of reviews to support online purchase decisions. For example, introducing a mechanism to allow an easier comparison between reviews would affect the process of making a purchase decision [83].

REFERENCES

[1] B. Pang and L. J. Lee, Opinion Mining and Sentiment Analysis. Hanover, MA: Now Publishers, 2008.

[2] P. Chatterjee, "Online Reviews: Do Consumers Use Them?," Social Science Research Network Working Paper Series, May/09/ 2006.

[3] P.-Y. Chen, S.-Y. Wu, and J. Yoon, "The impact of online recommendations and consumer feedback on sales," in ICIS 2004 Proceedings, p. 58, 2004.

[4] C. Dellarocas, N. Awad, and X. Zhang, "Exploring the value of online reviews to organizations: Implications for revenue forecasting and planning," in ICIS 2004 Proceedings, p. 30, 2004.

[5] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger, "Pulse: mining customer opinions from free text," in Advances Intelligent Data Analysis Madrid, Spain, 2005.

[6] Y. Chen and J. Xie, "Online Consumer Review: Word-of-Mouth as a New Element of Marketing Communication Mix," Manage. Sci., vol. 54, pp. 477-491, 2008.

[7] J. A. Chevalier and D. Mayzlin, "The effect of word of mouth on sales: Online book reviews," Journal of Marketing Research, vol. 43, pp. 345-354, 2006/08// 2006.

[8] Y. Liu, "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue," Journal of Marketing, vol. 70, pp. 74-89, 2006.

[9] B. Liu, Sentiment analysis and opinion mining. San Rafael, Calif.: Morgan & Claypool, 2012.

[10] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou, "Low-Quality Product Review Detection in Opinion Summarization," in Proceedings of (EMNLP-CoNLL), 2007, pp. 334-342.

[11] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti, "Automatically assessing review helpfulness," in the Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, 2006.

[12] A. Ghose and P. G. Ipeirotis, "Designing novel review ranking systems: predicting the usefulness and impact of reviews," in the Proceedings of the ninth international conference on Electronic commerce, Minneapolis, MN, USA, 2007.

[13] Z. Zhang and B. Varadarajan, "Utility scoring of product reviews," in the Proceedings of the 15th ACM international conference on Information and knowledge management, Arlington, Virginia, USA, 2006.

[14] A. Ghose and P. G. Ipeirotis, "Designing novel review ranking systems: predicting the usefulness and impact of reviews," in the Proceedings of the ninth international conference on Electronic commerce, Minneapolis, MN, USA, 2007.

[15] A. Ghose and P. G. Ipeirotis, "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics," Knowledge and Data Engineering, IEEE Transactions on, vol. 23, pp. 1498-1512, 2011.

[16] Y. Liu, X. Huang, A. An, and X. Yu, "Modeling and Predicting the Helpfulness of Online Reviews," in the Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, 2008.

[17] S. Huang, D. Shen, W. Feng, Y. Zhang, and C. Baudin, "Discovering clues for review quality from author's behaviors on e-commerce sites," in the Proceedings of the 11th International Conference on Electronic Commerce, Taipei, Taiwan, 2009.

[18] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi, "Exploiting social context for review quality prediction," in the Proceedings of the 19th international conference on World Wide Web, Raleigh, North Carolina, USA, 2010.

[19] X. Yu, Y. Liu, X. Huang, and A. An, "A quality-aware model for sales prediction using reviews," in the Proceedings of the 19th international conference on World wide web, Raleigh, North Carolina, USA, 2010.

[20] M. P. O'Mahony and B. Smyth, "Learning to recommend helpful hotel reviews," in the Proceedings of the third ACM conference on Recommender systems, New York, New York, USA, 2009.

[21] C. C. Chen and Y.-D. Tseng, "Quality evaluation of product reviews using an information quality framework," Decision Support Systems, vol. 50, pp. 755-768, 2011.

[22] R. Y. Wang and D. M. Strong, "Beyond accuracy: what data quality means to data consumers," J. Manage. Inf. Syst., vol. 12, pp. 5-33, 1996.

[23] O. Tsur and A. Rappoport, "RevRank: A Fully Unsupervised Algorithm for Selecting the Most Helpful Book Reviews," in Proceedings of the International AAAI Conference on Weblogs and Social Media 2009.

[24] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Texts," in Empirical Methods in Natural Language Processing Conference 2004, pp. 404-411.

[25] X. Wan and J. Xiao, "CollabRank: towards a collaborative approach to single-document keyphrase extraction," in the Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1, Manchester, United Kingdom, 2008.

[26] S. Moghaddam, M. Jamali, and M. Ester, "ETF: extended tensor factorization model for personalizing prediction of review helpfulness," in the Proceedings of the fifth ACM international conference on Web search and data mining, Seattle, Washington, USA, 2012.

[27] P. Tsaparas, A. Ntoulas, and E. Terzi, "Selecting a comprehensive set of reviews," in the Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, San Diego, California, USA, 2011.

[28] T. Lappas and D. Gunopulos, "Efficient confident search in large review corpora," in the Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part II, Barcelona, Spain, 2010.

[29] Q. Miao, Q. Li, and R. Dai, "AMAZING: A sentiment mining and retrieval system," Expert Systems with Applications, vol. 36, pp. 7192-7198, 4// 2009.

[30] G. Wu, D. Greene, and P. Cunningham, "Merging multiple criteria to identify suspicious reviews," in Proceedings of the fourth ACM conference on Recommender systems, 2010, pp. 241-244.

[31] R. Y. Lau, W. Zhang, Y. Xia, and D. Song, "Multi-facets quality assessment of online opinionated expressions," in Web Information Systems Engineering–WISE 2010 Workshops, 2011, pp. 212-225.

[32] H.-J. Min and J. C. Park, "Identifying helpful reviews based on customer's mentions about experiences," Expert Systems with Applications, vol. 39, pp. 11830-11838, 11/1/ 2012.

[33] Y. Liu, J. Jin, P. Ji, J. A. Harding, and R. Y. K. Fung, "Identifying helpful online reviews: A product designer's perspective," Computer-Aided Design, vol. 45, pp. 180-194, 2// 2013.

[34] Y.-C. Zeng and S.-H. Wu, "Modeling the Helpful Opinion Mining of Online Consumer Reviews as a Classification Problem," in Sixth International Joint Conference on Natural Language Processing, 2014, p. 29.

[35] L. Connors, S. M. Mudambi, and D. Schuff, "Is It the Review or the Reviewer? a Multi-Method Approach to Determine the Antecedents of Online Review Helpfulness," in System Sciences (HICSS), 2011 44th Hawaii International Conference on, 2011, pp. 1-10.

[36] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, and L. Lee, "How opinions are received by online communities: a case study on amazon.com helpfulness votes," in the Proceedings of the 18th international conference on World wide web, Madrid, Spain, 2009.

[37] B. Pang and L. Lee, "Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales," in the Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Ann Arbor, Michigan, 2005.

[38] Y. Attali and J. Burstein, "Automated essay scoring with e-rater v.2," Journal of Technology, Learning, and Assessment, vol. 4, 2006/02// 2006.

[39] J. Nitin and B. Liu, "Product Review Analysis," ed: Technical Report, UIC, 2007.

[40] M. Weimer, I. Gurevych, and M. Mühlhäuser, "Automatically assessing the post quality in online discussions on software," in Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, 2007, pp. 125-128..

[41] M. Weimer and I. Gurevych, "Predicting the Perceived Quality of Web Forum Posts," in the Proceeding of Recent Advances in Natural Language Processing Conference (RANLP), 2007.

[42] J. Jeon, W. B. Croft, J. H. Lee, and S. Park, "A framework to predict the quality of answers with non-textual features," in the Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, Washington, USA, 2006.

[43] L. Hoang, J.-T. Lee, Y.-I. Song, and H.-C. Rim, "A model for evaluating the quality of user-created documents," in the Proceedings of the 4th Asia information retrieval conference on Information retrieval technology, Harbin, China, 2008.

[44] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in the Proceedings of the 2008 International Conference on Web Search and Data Mining, Palo Alto, California, USA, 2008.

[45] J. Bian, Y. Liu, D. Zhou, E. Agichtein, and H. Zha, "Learning to recognize reliable users and content in social media with coupled mutual reinforcement," in the Proceedings of the 18th international conference on World wide web, Madrid, Spain, 2009.

[46] Y. Y. Hao, Y. J. Li, and P. Zou, "Why Some Online Product Reviews Have No Usefulness Rating," in the Proceeding of the Pacific Asia Conference of Information Systems 2009.

[47] T. Khopkar, X. Li, and P. Resnick, "Self-selection, slipping, salvaging, slacking, and stoning: the impacts of negative feedback at eBay," in the Proceedings of the 6th ACM conference on Electronic commerce, Vancouver, BC, Canada, 2005.

[48] P. Resnick, R. Zeckhauser, J. Swanson, and K. Lockwood, "The value of reputation on eBay: A controlled experiment," Experimental Economics, vol. 9, pp. 79-101, 2006/06/01 2006.

[49] P.-Y. Chen, S. Dhanasobhon, and M. Smith, "All Reviews are Not Created Equal: The Disaggregate Impact of Reviews and Reviewers at Amazon.Com," Social Science Research Network Working Paper Series, July/20/ 2006.

[50] S. Huang, D. Shen, W. Feng, C. Baudin, and Y. Zhang, "Improving product review search experiences on general search engines," in the Proceedings of the 11th International Conference on Electronic Commerce, Taipei, Taiwan, 2009.

[51] M. A. Suryanto, E. P. Lim, A. Sun, and R. H. L. Chiang, "Quality-aware collaborative question answering: methods and evaluation," in the Proceedings of the Second ACM International Conference on Web Search and Data Mining, Barcelona, Spain, 2009.

[52] S. M. Mudambi, "What makes a helpful online review? A study of customer reviews on amazon.com," MIS Quart Manage Inf Syst MIS Quarterly: Management Information Systems, vol. 34, pp. 185-200, 2010.

[53] N. Korfiatis, E. García-Bariocanal, and S. Sánchez-Alonso, "Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content," Electronic Commerce Research and Applications, vol. 11, pp. 205-217, 5// 2012.

[54] Q. Cao, W. Duan, and Q. Gan, "Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach," Decis. Support Syst., vol. 50, pp. 511-521, 2011.

[55] M. Siering and J. Muntermann, "What Drives the Helpfulness of Online Product Reviews? From Stars to Facts and Emotions," in Wirtschaftsinformatik, 2013, p. 7.

[56] Y. Dezhi, B. Samuel, and Z. Han, "Dreading and Ranting: The Distinct Effects of Anxiety and Anger in Online Seller Reviews, " in the Poceeding In International Conference on Information Systems, 2011.

[57] P. F. Wu, H. V. d. Heijden, and N. Korfiatis, "The Influences of Negativity and Review Quality on the Helpfulness of Online Reviews," in the International Conference on Information Systems, 2011.

[58] Z. Yang and X. Fang, "Online service quality dimensions and their relationships with satisfaction: A content analysis of customer reviews of securities brokerage services," International Journal of Service Industry Management, vol. 15, pp. 302-326, 2004.

[59] R. M. Schindler and B. Bickart, "Perceived helpfulness of online consumer reviews: The role of message content and style," Journal of Consumer Behaviour, vol. 11, pp. 234-243, 2012.

[60] P. Racherla and W. Friske, "Perceived 'usefulness' of online consumer reviews: An exploratory investigation across three services categories," Electron. Commer. Rec. Appl., vol. 11, pp. 548-559, 2012.

[61] J. Otterbacher, "'Helpfulness' in online communities: a measure of message quality," in the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, MA, USA, 2009.

[62] A. E. Schlosser, "Can including pros and cons increase the helpfulness and persuasiveness of online reviews? The interactive effects of ratings and arguments," Journal of Consumer Psychology, vol. 21, pp. 226-239, 7// 2011.

[63] J. Tang, H. Gao, X. Hu, and H. Liu, "Context-aware review helpfulness rating prediction," in the Proceedings of the 7th ACM conference on Recommender systems, Hong Kong, China, 2013.

[64] C. Dellarocas, "The digitization of word of mouth: Promise and challenges of online feedback mechanisms, " Management science, vol. 49, pp. 1407-1424, 2003.

[65] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins, "Propagation of trust and distrust, " in Proceedings of the 13th international conference on World Wide Web, 2004, pp. 403-412.

[66] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina, "The Eigentrust algorithm for reputation management in P2P networks, " in the Proceedings of the 12th international conference on World Wide Web, Budapest, Hungary, 200

[67] S. Kurohashi, K. Inui, and Y. Kato, eds., Workshop on Information Credibility on the Web, 2007.

[68] P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman, "Reputation systems," (CACM), vol. 43, pp. 45–48, (ISSN 0001-0782), 2000.

[69] S. Krishnamoorthy, "Linguistic Features for Review Helpfulness Prediction," Expert Systems with Applications, 2015.

[70] G. R. Semin and K. Fiedler, "The linguistic category model, its bases, applications and range," European review of social psychology, vol. 2, pp. 1-30, 1991.

[71] S. Lee and J. Y. Choeh, "Predicting the helpfulness of online reviews using multilayer perceptron neural networks," Expert Systems with Applications, vol. 41, pp. 3041-3046, 2014.

[72] Z. Zhang, Y. Ma, G. Chen, and Q. Wei, "Extending associative classifier to detect helpful online reviews with uncertain classes, " 2015.

[73] T. L. Ngo-Ye and A. P. Sinha, "The influence of reviewer engagement characteristics on online review helpfulness: A text regression model, " Decision Support Systems, vol. 61, pp. 47-58, 2014.

[74] Z. Liu and S. Park, "What makes a useful online review? Implication for travel product websites, " Tourism Management, vol. 47, pp. 140-151, 2015.

[75] S. Quaschning, M. Pandelaere, and I. Vermeir, "When Consistency Matters: The Effect of Valence Consistency on Review Helpfulness, " Journal of Computer‐Mediated Communication, vol. 20, pp. 136-152, 2015.

[76] K. Carlson and A. Guha, "The ratings paradox: Why we prefer reading negative reviews, but then subsequently rate these reviews as less useful," Advances in Consumer Research, vol. 37, 2010.

[77] Y. Pan and J. Q. Zhang, "Born unequal: a study of the helpfulness of user-generated product reviews," Journal of Retailing, vol. 87, pp. 598-612, 2011.

[78] M. Scholz and V. Dorner, "The Recipe for the Perfect Review?," Business & Information Systems Engineering, vol. 5, pp. 141-151, 2013/06/01 2013.

[79] L. Martin and P. Pu, "Prediction of helpful reviews using emotions extraction," in Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.

[80] M. Mertz, N. Korfiatis, and R. V. Zicari, "Using Dependency Bigrams and Discourse Connectives for Predicting the Helpfulness of Online Reviews," in E-Commerce and Web Technologies, ed: Springer, 2014, pp. 146-152.

[81] S. M. Mudambi, D. Schuff, and Z. Zhang, "Why Aren't the Stars Aligned? An Analysis of Online Review Content and Star Ratings," in System Sciences (HICSS), 2014 47th Hawaii International Conference on, 2014, pp. 3139-3147.

[82] A. H. Huang, K. Chen, D. C. Yen, and T. P. Tran, "A study of factors that contribute to online review helpfulness," Computers in Human Behavior, vol. 48, pp. 17-27, 2015.

[83] G. Lackermair, D. Kailer, and K. Kanmaz, "Importance of Online Product Reviews from a Consumer's Perspective," 2013.