International Journal of Semantic Computing
© World Scientific Publishing Company

EVALUATING ONLINE PRODUCTS USING TEXT MINING: A RELIABLE EVIDENCE-BASED APPROACH

HAIPING XU*

*Computer and Information Science Department, University of Massachusetts Dartmouth, 285 Old Westport Rd,
Dartmouth, MA 02747, USA
hxu@umassd.edu
http://www.cis.umassd.edu/~hxu*

RAN WEI

*Computer and Information Science Department, University of Massachusetts Dartmouth, 285 Old Westport Rd,
Dartmouth, MA 02747, USA
rwei@umassd.edu*

RICHARD DEGROOF

*Computer and Information Science Department, University of Massachusetts Dartmouth, 285 Old Westport Rd,
Dartmouth, MA 02747, USA
rdegroof@umassd.edu*

JOSHUA CARBERRY

*Computer and Information Science Department, University of Massachusetts Dartmouth, 285 Old Westport Rd,
Dartmouth, MA 02747, USA
jcarberry@umassd.edu*

Received (12 March 2022)
Revised (15 June 2022)
Accepted (02 July 2022)

To address the uncertainty about the quality of online merchandise, e-commerce sites often provide product review ranking services to help customers make purchasing decisions. Such services can be very useful, but they are not necessarily reliable when the ranking results are based on ratings without considering their reliability. In this paper, we propose a reliable evidence-based approach to online product evaluation by using text mining to analyze product reviews while taking into account the reliability of each review. We parse the product reviews and classify the opinion orientations for each recognized product feature as positive or negative. Then, we weight the classified opinion orientations by their reliability and use them as independent evidence to calculate the belief values of the product using Dempster-Shafer (D-S) theory. Based on the belief values of a list of similar products, we can calculate their product effectiveness and cost-effectiveness values for product ranking. The case studies show that our approach can greatly help customers make better decisions when choosing the right online products.

Keywords: E-commerce; product reviews; text mining; cost effectiveness; reasoning under uncertainty; Dempster-Shafer (D-S) theory.

*Corresponding author: Haiping Xu, Computer and Information Science Department, University of Massachusetts Dartmouth, Dartmouth, MA 02747, USA, Email: hxu@umassd.edu

1. Introduction

E-commerce technology provides users with the great benefits of online shopping such as incredible convenience, helpful product reviews, reasonable prices, and a wide selection of products. Nowadays, more and more people prefer to purchase products online rather than shopping directly in a physical store. However, due to the inherent nature and complexity of e-commerce environment (e.g., the wide variety of brands and the large number of similar products available online), it is often difficult for customers to decide which products they should choose without having the necessary technical knowledge about the products they want to purchase. To help customers choose products that meet their requirements, an accurate evaluation of online products becomes increasingly important. Online stores, such as Amazon, have attempted to develop suitable and effective mechanisms for product evaluation [1]. However, as noted in previous studies, such mechanisms are typically not as well-employed as expected because customers rarely view online comments beyond the first two review pages due to consumer attention constraints [2, 3]. Thus, while customers can reasonably be expected to manually assess a handful of user reviews, they do not have the attention span or time to draw meaningful conclusions from the dozens or hundreds of reviews that are often left on popular products. Furthermore, since average ratings have become the de facto aggregate review scores used by many websites, it is much easier for customers to check average ratings than to read all product reviews on multiple web pages [4].

We realize that most product ranking services that use rating mechanisms, such as average star rating (ASR), are not always reliable. This is because customer ratings are usually based on customers' subjective tastes of products and are prone to bias [4]. In addition, the incentive of leaving ratings due to "bragging and moaning" may also lead to a bi-modal and non-normal distribution of product ratings, which makes the average product score a misleading recommendation of the product's true quality [5]. In this paper, instead of using customer ratings to rank products, we focus on the text of product reviews and analyze them using text mining to identify substantial evidence to support reliable online product evaluation. To make our approach more trustworthy, we further consider the reliability of each review. We notice that users tend to ignore useful information related to a product review, such as the helpfulness of the review rated by other users and the qualification of the reviewer, which can be used as evidential knowledge for evaluating the reliability of the product review. To exploit this "hidden" knowledge, we define multiple attributes related to product reviews and set evaluation criteria for each attribute, quantified using certain scales.

In our approach, we investigate whether a product is a favorable one worth buying or an unfavorable one not worth buying. To draw meaningful conclusions, we analyze all review comments of a product using the part-of-speech (POS) tagging model and classify user opinion orientations on a selected set of product features as positive or negative. We first compute the weighted average counts of positive and negative opinion orientation on a particular product feature using the reliability of each review. The weighted average counts of positive and negative opinions on the product feature are then considered as conflicting evidence that can be combined using Dempster-Shafer (D-S) theory [6] to support statements about product favorability. Note that D-S theory is a mathematical theory of evidence that is a powerful tool for supporting reasoning under uncertainty [7]. Using Dempster's combination rules, multiple independent pieces of evidence can be combined to derive a high-level degree of belief for specific working assumptions, namely hypotheses.

In the following steps of our method, we consider all selected product features and number of reviews as independent evidence and further combine them to calculate the product effectiveness and cost-effectiveness values. By ranking various online products sold by different vendors based on their cost-effectiveness values, our approach can greatly assist e-commerce customers in making decisions about choosing the most cost-effective products. This work significantly extends our previous proposed framework for assessing the quality of online products using D-S theory [8]. In our previous work, we focused on product's review scores, categorized them as positive or negative, and used them as independent evidence to evaluate the product. To overcome the drawbacks of using review values such as ASR scores for product ranking, in this paper we further analyze the text of the review comments and derive more meaningful evidence for online product evaluations. To provide reliable online product evaluations, we propose a systematic approach to extract fine-grained sentiments of the review texts and combine them with properties of the reviews using D-S theory. While D-S theory of evidence is a classical model developed in the 1970s, the novelty of our proposed approach lies in the use of contradictory sentiments in product reviews in conjunction with their reliability. Our approach integrates conflicting evidence and positive / negative sentiment into a single value for product effectiveness and cost-effectiveness. These proposed metrics not only provide a more balanced and objective measure for evaluating online products, but are also well scalable, as any additional evidence can be combined with existing calculations without the need for a complete re-evaluation.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 reviews D-S theory. Section 4 presents a conceptual model for online product evaluation and a framework for our reliable evidence-based approach. Section 5 discusses how text mining can be used to identify product features and user opinion orientations on selected product features. Section 6 provides details of the product evaluation process. Section 7 presents two case studies and their analysis results. Section 8 concludes the paper and mentions future work.

2. Related Work

To manage uncertainty in many domains, researchers have employed various methods to support reasoning under uncertainty. Foundational approaches include imprecise probability, which models uncertain probabilities as values within intervals defined by lower and upper bounds [9], and possibility theory, which initially uses fuzzy sets and fuzzy logic to express uncertainty [10]. Researchers have also extended Markov models by incorporating first-order logic with probability to represent uncertain situations [11]. In our evidence-based approach, we adopt D-S theory, which is a theory of evidence that provides a mathematical representation of uncertainty. Due to its ability to handle conflicting information, D-S theory has been widely used for information fusion in many different domains. Dong *et al.* proposed a practical shill detection mechanism in online auctions using D-S theory of evidence [12]. The approach takes into account multiple pieces of evidence from different information layers, detects shilling behaviors, and assists decision making on shill bidders. Panigrahi *et al.* developed a fraud detection system for mobile communication networks [13]. They utilized D-S theory to combine multiple pieces of evidence captured from a rule-based component and computed an overall suspicion score to help users filter suspicious incoming calls. Zhang *et al.* demonstrated the effectiveness of D-S theory for the service supplier selection problem [14]. D-S theory was shown to provide valuable insights to the decision-maker despite the high-uncertainty and low-information nature of supplier selection, a problem with broad and sometimes unclear criteria. Li and Wei developed an approach to incorporating D-S theory to assist decision making during disaster and emergency situations [15]. D-S theory was used to enhance existing decision-making methods based on probabilistic linguistic term sets. The aim of the enhancements was to reduce information loss and assist decision makers in highly uncertain environment of disasters. Yang and Mandan proposed an evidential reasoning approach that can be used to solve uncertain decision problems with quantitative and qualitative properties [16]. They proposed an alternative way to handle hybrid multi-attribute decision-making problems with uncertainty. In contrast to these methods, in this paper we develop a formal cost-effectiveness analysis model for online products based on D-S theory. Our approach combines multiple pieces of evidence extracted from the reviews of a product and evaluates the quality of the product by calculating its effectiveness value.

Previous research work on product analysis is summarized as follows. Cho and Kim developed a product taxonomy for collaborative recommendation in e-commerce [17]. In their approach, they used web usage mining techniques to improve performance and recommendation quality. Sarwar *et al.* proposed an analytic-recommendation algorithm that can generate useful recommendations to customers [18]. They used traditional methods such as data mining and dimensionality reduction techniques to deal with large-scale purchase and preference data. Zhang and Feng developed a pricing model for identical products sold in gray markets [19]. They showed that the price gap between two separate markets could affect consumer demand and positively influence gray market sales with increasing profits. The above methods are based on analysis of large data sets and are useful for product recommendations; as such, they complement our research efforts to analyze the review comments associated with specific online products.

There is also some existing work on detecting review spammers and filtering other low-quality reviews by analyzing product reviews. Wang *et al.* proposed a new concept of a heterogeneous review graph to capture the relationships between reviewers and the stores that the reviewers have reviewed [20]. They explored how interactions between nodes can be used to identify suspicious reviewers. Li *et al.* exploited machine learning methods to identify spam reviews [21]. By using supervised learning methods and analyzing the effect of various product features, their method can perform better than existing heuristic-based methods. Wang *et al.* employed an extended LDA (Latent Dirichlet Allocation) model to detect fraudulent reviews [22]. They adapted the traditional LDA by categorizing the reviews into genuine or fraudulent classes, which are further clustered based on similarities and levels of suspicion. Mukherjee *et al.* studied spam detection in a collaborative setting to identify fake reviewer groups [23]. They proposed a new relation-based model, called *GSRank*, which considers relationships among groups, individual reviewers, and the products reviewed for detection of spammer groups. Heng *et al.* used a topic modeling approach to determine metrics related to the quality of reviews [24]. Their approach enables rating the subjectivity of reviews, which is an important factor in determining the potential reliability of reviews. Note that the methods described above provide useful means for assessing the validity and usefulness of reviews; therefore, they have the potential to be integrated into our evidence-based approach to improve accuracy and validity.

More recently, several research efforts have focused on feature identification using unsupervised learning methods (e.g., LDA) in mining online e-commerce review comments for sentiment analysis. Zhu *et al.* proposed TipSelector, an unsupervised algorithm that provides high quality hints without annotated training data [25]. They reported the tuning of the *K* parameter, the number of latent topics in the LDA model, to derive their results. Lau *et al.* designed a new social analytics approach that processes archived consumer reviews on social media sites for fine-grained extraction of market intelligence [26]. They proposed a social analytics methodology that used a novel LDA-based semi-supervised fuzzy product ontology mining algorithm for aspect-oriented sentiment analysis. Castillo *et al.* developed a method to generate recommendations based on inferred user preferences [27]. They directly used existing product metadata and user rating patterns to determine the relationship between product attributes and personal preferences. Luo *et al.* used existing lexical databases and word embeddings to extract prominent review aspects [28]. Using the embeddings and existing knowledge in the database, reviews were processed into potential review aspects, which were then charted and clustered to produce a list of prominent, non-overlapping review aspects. The above approaches focus on extracting product features from product reviews; whereas our approach evaluates product quality by analyzing conflicting customer opinions based on a set of manually determined product features. Therefore, the above approaches can be very useful for us to improve our methodology by automatically extracting product features when dealing with a large number of various online products.

Much attention has been given to the sentiment analysis of product reviews. Haque *et al.* used supervised learning to predict the sentiment polarity of Amazon reviews [29]. Unlabeled review data were obtained from different sources and labeled using semi-supervised pool-based active learning. Various supervised learning methods were used on the labeled data and showed effectiveness in predicting the polarity of reviews. Shrestha and Nasoz used a recurrent neural network with gated recurrent units to predict the sentiment of Amazon reviews based on text and product information [30]. They used the predicted sentiment to determine the star rating of the reviews. Xu *et al.* focused on natural language processing to read and comprehend reviews while analyzing sentiment [31]. They used a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model that was trained to comprehend review text and can be used to extract and analyze the sentiment of product reviews. While these methods explore the sentiment of reviews, they did not attempt to use it to draw meaningful conclusions about the reviewed products. There has also been some previous work to develop a ranking system for Amazon products based on customer reviews. Zhang *et al.* measured text-derived sentiment and developed a graph-based method for ranking products [32]. The graph reflects the relative quality of products, which can be ranked by mining the graph using an algorithm similar to page rank. Ghose and Ipeirotis studied the impacts of product reviews on sales in terms of informativeness and influence [33]. In their approach, they developed a regression model to measure these attributes and provided a review ranking to predict the sales performance of products. Zha *et al.* considered sentiment in a probabilistic model of product ranking based on review comments [34]. They formulated an overall rating generated by a Gaussian distribution whose mean is calculated from weighted sentiment values. Recently, Xu *et al.* proposed a novel feature-based sentence model (FSM) that introduces a latent layer, called the feature layer, between review sentences and words [35]. In their approach, by performing sentiment analysis on each review sentence and deriving a weighted feature preference vector for the review, a review summary of a product can be developed that provides the user with the most desirable product features. While the above approaches consider sentiment-based product rankings, they differ from our evidence-based approach because none of them address the issue of how to aggregate conflicting evidence to support the fusion of uncertain information to infer product quality.

In summary, e-commerce applications using D-S theory to support reasoning under uncertainty are rare. Since product reviews were written by individuals who may have different opinions about product quality, there is a pressing need to apply suitable reasoning mechanisms, such as D-S theory, to support the aggregation of different opinions. Existing studies on product ranking are usually not evidence-based but rely on score ratings or conventional data mining methods. Our proposed approach fills the research gap by using feature keywords and sentiments augmented with synonyms to derive more meaningful product quality metrics. In addition, our approach is scalable because incorporating new features or comments requires only a simple recalculation of the effectiveness value, folding in the new evidence.

3. Dempster-Shafer Theory of Evidence

D-S theory is an evidence-based probabilistic reasoning approach developed to address the uncertainty and incompleteness of the available information [6], [7]. Let Θ be a finite set of mutually exclusive possible hypotheses, called a *frame of discernment*. For example, when we consider the product evaluation domain, whether a product is *favorable* or *unfavorable* for purchase depends on the nature of the evaluation properties and the quantified values of the review evidence. Thus, the frame of discernment for a product can be defined as Θ = {*favorable*, *unfavorable*}. The power set of Θ that contains all subsets of Θ is defined as *P*(Θ) = {∅, {*favorable*}, {*unfavorable*}, Θ}.

In D-S theory, a belief mass (also called mass or mass value) is assigned to each element of the power set *P*(Θ) in the interval between 0 and 1. Thus, the basic mass assignment (*BMA*) function *m* is defined as *m*: *P*(Θ) → [0, 1], which satisfies the following two requirements:

$$m(\varnothing) = 0 \tag{1}$$

$$\sum_{H \in P(\Theta)} m(H) = 1 \tag{2}$$

In Eq. (1), the mass of the empty set ∅ represents the measure of none state, thus it is defined as 0. Eq. (2) indicates that the sum of masses of the elements in the power set is equal to 1. For example, in our product review example, since *m*(∅) = 0, we have *m*({*favorable*}) + *m*({*unfavorable*}) + *m*(Θ) = 1. Note that *m*(Θ) represents the mass of conflicting states (i.e., in our example a hypothesis says that a product is both favorable and unfavorable), and thus it can be interpreted as a measure of uncertainty. For clarity, in the rest of the paper, we use the notation *m*(*U*) to denote *m*(Θ), where *U* represents uncertainty.

For a set of states (or a hypothesis) *H*, another important function is called the *belief* function, which is defined as the sum of the masses of all subsets of *H*.

$$belief(H) = \sum_{G \subseteq H} m(G) \tag{3}$$

Intuitively, any portion of a belief committed to the hypothesis implied by hypothesis *H* must also be committed to hypothesis *H*. Thus, to obtain the total belief in *H*, the quantities *m*(*G*) for ∀*G* ⊆ *H* must be added to *m*(*H*). In our example, we have two hypotheses, namely 1) the product is favorable; and 2) the product is unfavorable, both of which have no proper subset except for ∅. Thus, according to Eq. (3), the belief values *belief*({*favorable*}) = *m*({*favorable*}) and *belief*({*unfavorable*}) = *m*({*unfavorable*}).

The Dempster's rule of combination is a key concept of the original idea of D-S theory. Given two mass values m_a and m_b for hypothesis *H*, the combination rule computes the *joint mass* for the two pieces of evidence *a* and *b* under the same hypothesis *H*, which can be calculated as in Eqs. (4) and (5).

$$m_{a,b}(\varnothing) = 0 \tag{4}$$

$$m_{a,b}(H) = m_a(H) \oplus m_b(H) = \frac{1}{1-C} \sum_{H_1 \cap H_2 = H} m_a(H_1) m_b(H_2) \tag{5}$$

where $C = \sum_{H_1 \cap H_2 = \varnothing} m_a(H_1) m_b(H_2)$ and $H \neq \varnothing$.

Eq. (4) says that the combined mass for the empty set $\varnothing$ is zero. In Eq. (5), C represents the measure of the amount of conflict between the two mass sets. This is determined by the sum of the masses of any pair of sets H_1 and H_2, where H_1 and H_2 are disjointed subsets of Θ. Therefore, $(1 - C)$ can be used as a normalization factor, which serves to ignore any conflict between disjointed pairs of states.

Traditionally, Dempster's rule has been interpreted as an operator that fuse separate argument beliefs from independent sources into a single belief [36]. Although some examples with conflicting evidence may show counterintuitive results when using Dempster's rule, many researchers have provided reasonable explanations for this phenomenon [37]-[39]. In our approach, we take into account the reliability of the evidence sources and introduce weighted mass values for uncertainty. Since at least one of the evident sources must be *unreliable* in the case of conflicting evidence, by applying the reliability of the evidence sources, D-S theory yields reasonable results, as demonstrated in our previous studies [8], [12].

4. A Reliable Evidence-Based Approach

4.1. *A Conceptual Model*

To help customers verify the quality of an online product, we introduce a formal cost-effectiveness analysis model for product evaluation in e-commerce. In our approach, the quality of an online product is reflected by its effectiveness value based on the information collected from an electronic marketplace. We now define the concepts of product effectiveness and cost effectiveness as follows.

Definition 1: The *product effectiveness* is defined as a function E: $P \rightarrow [0, 1]$, where P is a set of product alternatives, each of which is mapped to an effectiveness value between 0 and 1. The product effectiveness of product alternative $p \in P$ quantifies the product quality of p based on the belief of the product state whether it is worth buying.

Definition 2: The *cost effectiveness* of a product is defined as a function CE: $P \rightarrow R_0^+$, where P is a set of product alternatives, each of which is mapped to a cost-effectiveness value that is a nonnegative real number. The *cost effectiveness* of a product alternative $p \in P$ quantifies the product effectiveness and the product cost of p.

E-commerce websites like Amazon usually offer flexible platforms that contain a large amount of useful information related to product quality. For example, at Amazon website, not only is a customer who has purchased a product online allowed to provide a review star rating and review comments for the product, but the review can also be further rated by other customers. For popular products, the review information is spread over multiple pages, and most customers usually ignore these pages except for the first few pages or pages containing very negative reviews. To support automated analysis of this useful information for making decisions about online purchases, we take all reviews and their associated properties as evidence supporting the favorability or unfavorability of a product and derive the cost-effectiveness analysis model. The conceptual model of cost-effectiveness analysis in e-commerce can be formally defined as a 3-tuple (P, Bel, $mCost$), where

(1) $P = \{p_1, p_2, \ldots, p_n\}$ is a set of product alternatives that need to be evaluated and ranked, which must have similar functionality and be in the same price range;
(2) Bel: $P \rightarrow [0, 1]$ is a belief function used in our model that maps each product alternative to a degree of belief that quantifies whether the product is worth buying;
(3) $mCost$: $P \rightarrow R_0^+$ is a cost function that maps a product alternative to its minimal price, defined as a nonnegative real number. Note that for a particular product $p \in P$, we can use $Bel(p)$ and $mCost(p)$ to calculate its cost-effectiveness value, which can then be used to rank the product alternatives in P.

To evaluate each product $p \in P$, we define p as a 6-tuple (REV, FE, $PROP$, Rel, EV, M), where

(1) $REV = \{r_1, r_2, \ldots, r_n\}$ is a set of n product reviews about product p, provided by different reviewers;
(2) $FE = \{f_1, f_2, \ldots, f_k\}$ is a set of k selected product features of product type P for evaluating the effectiveness of product p;
(3) $PROP = \{pr_1, pr_2, \ldots, pr_i\}$ is a set of i review properties that help to calculate the reliability of each review;
(4) Rel: $REV \rightarrow [0, 1]$ is a reliability function of product reviews, which indicates the importance and accuracy of each review;
(5) $EV = \{ev_1, ev_2, \ldots, ev_l\}$ is a set of evidence used to justify a product favorable or unfavorable, where $l = 2*k+1$;
(6) $M = \{m: EV \rightarrow [0, 1]\}$ is a set of mass assignment functions that quantify and assess each piece of evidence supporting the product as favorable or unfavorable.

In our analytical model, we consider product features as evidence and identify positive and negative opinion orientations from product reviews. For a specific product feature, such as "Sound" of a speaker product, we go through each review comment and find opinion words for the feature of "Sound" to determine whether the reviewer's opinion is positive or negative. Note that for multiple opinion words with the same opinion orientation, we count them only once because they represent the same piece of evidence from the reviewer. Since the total number of reviews for a product represents a strong indication whether the product is a popular product, we consider the number of reviews as a special product feature that can be used as independent evidence in combination with other evidence (i.e., positive and negative opinion orientations from product reviews). Let the number of selected product features for product type P be k. The total number of evidence l is equal to $2*k + 1$, including k positive-negative pairs of product features and the number of reviews. All pieces of evidence are quantified using the mass assignment functions to calculate their mass values and combined using the rule of Dempster to derive the belief values of the product. For a set of product alternatives $P = \{p_1, p_2, \ldots, p_n\}$, the function *Bel*: $P \rightarrow [0, 1]$ maps each product to a value that quantifies whether the product is worth buying. By further considering the minimal cost for each product in *P*, we can calculate the cost-effectiveness values of all products in *P* and rank them accordingly.

4.2. *A Framework for Evaluating an Online Product*

Figure 1 shows a framework for evaluating an online product using text mining. Let *P* be a product type representing a set of similar online products. To evaluate the quality of an online product $p \in P$, we first need to identify the major product features and opinion words for *P*. For example, "*Installation*" is one of the important features of the product type "*Audio/Video Receiver*," and the words associated with this feature are "*installation*," "*setup*," "*connection*" and so on. On the other hand, "*easy*" and "*difficult*" are positive and negative opinion words, respectively, related to the "*Installation*" feature of this product type. This information must first be captured and stored as a dataset in the *Product Features* and *Opinion Words* database.

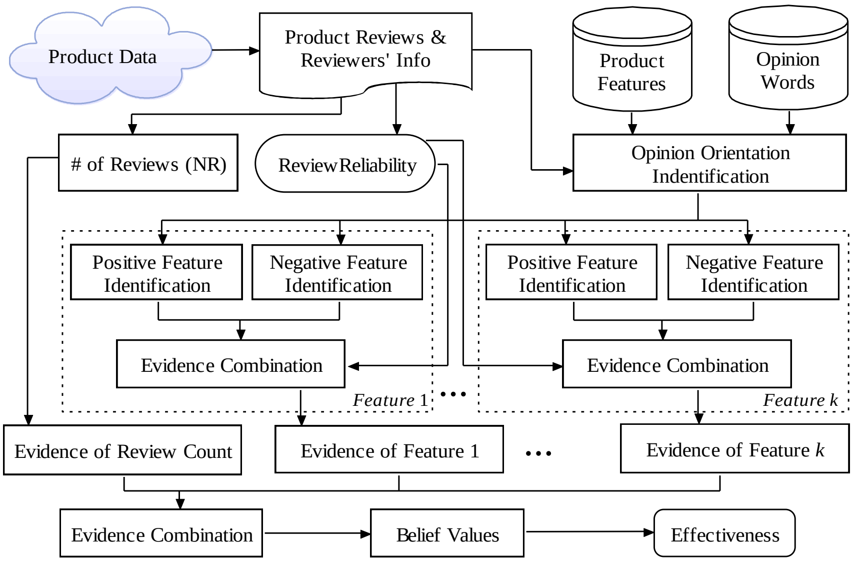

Fig. 1. A framework for evaluating an online product.

To evaluate a particular product *p*, we download all its product reviews as well as information about the reviewers. For each review $r \in REV$, once the relevant review properties are extracted, its reliability can be calculated. The reliability of each review *r* will be used to weight the opinion orientations in *r* when we count the number of positive and negative opinions from all reviews in set *REV*.

For each product feature *j*, we analyze the text of each review *r* to determine its opinion orientation, which should be positive to support that the product is favorable (denoted as *Positive Feature Identification* in Fig.1) or negative to support that the product is unfavorable (denoted as *Negative Feature Identification* in Fig. 1). Then, we combine them as conflicting evidence using Dempster's rule.

When combining the conflicting pieces of evidence, we must consider the reliability of each review and derive reliable mass values for the product feature. Once all mass values of the product features have been calculated, they can be used as independent evidence to justify the effectiveness of the product. As shown in Fig. 1, the total number of reviews for the product is considered a special feature when the pieces of evidence are combined. We calculate the *belief values* of the product as the output of the evidence combination process and further derive its product *effectiveness* value, where the detailed calculations are defined in Section 6.3.

5. Mining Online Product Reviews

5.1. *Part-of-Speech (POS) Tagging*

Product features are usually nouns or noun phrases appearing in multiple product reviews, while opinion words are usually adjectives in review sentences. Therefore, part-of-speech (POS) tagging is crucial in our text mining approach. Similar to previous work on mining and summarizing customer reviews [40], we used a toolkit called OpenNLP [41] to parse each review into sentences and generate POS tags for each word to indicate whether the word is a noun, verb, adjective, or other POS. Note that OpenNLP is a machine learning based toolkit for processing natural language text. Here we use an example to illustrate how a POS tagger can parse a sentence. Let the sentence be

"I love this camera, it is awesome."

Using the OpenNLP toolkit, the parser outputs the sentence with the following POS tags:

I_PRP love_VB this_DT camera_NN ,_, it_DT is_VB awesome_JJ ._.

In the above tagged sentence, the tags *PRP*, *VB*, *DT*, *NN* and *JJ* stand for "*Personal pronoun*," "*Verb*," "*Determiner*," "*Noun*," "*Adjective*," respectively. After detecting all sentences in a review and parsing each sentence with POS tags, we can save each sentence along with the POS tag information into a review database, which can be used to identify product features and opinion words, as described in the following sections.

5.2. *Product Feature Selection*

To identify product features, we search frequently used nouns and noun phrases from the review database and manually determine whether they are the appropriate words/phrases for the product features of a particular product. Note that a product feature can be expressed in different ways. For example, when customers talk about the installation of some equipment, they may use words or phrases such as "*setup*," "*connection*" and "*installation*." Thus, the product feature of "*Installation*" should contain all such major terms, which we call the *product feature set*. In determining the product feature set, we may also refer to existing product review publications such as *Consumer Reports*, which include feature lists of relevant product attributes with their reviews. Most features can be easily identified. For example, in the following sentence:

"The A/V receiver's setup is pretty easy."

The reviewer is satisfied with the installation of the *A/V receiver*, so "setup" is the feature that the reviewer comments on. However, due to the complexity and difficulty of natural language understanding, product features may also be implicitly mentioned, making them hard to be automatically captured. For example, we may have the following review sentence about a tablet:

"While light, it is not easily fit in pockets."

In this sentence, the customer implicitly refers to "*Tablet Size*" as a product feature, but the word "size" does not appear in this sentence. To simplify matters, in this paper we focus on identifying features that appear *explicitly* as nouns or phrases in product reviews. Identification of *implicit* features in reviews is envisioned as a more ambitious direction for our future research.

5.3. *Opinion Words Mining*

Customers may use various adjectives to express their opinions about specific product features. Such adjective words are used to tell whether the product is a favorable or an unfavorable product with respect to its product features. We collect and store them as a set called *opinion word set*, where each opinion word is labeled as "*positive*" or "*negative*" for the relevant product feature. Opinion words are usually adjectives that express positive, negative or neutral sentiments. Words that encode a desirable state (e.g., *great, nice, wonderful*) have a positive orientation; while words that represent an undesirable state (e.g., *disappointing, useless*) have a negative orientation. Note that opinion words representing neutral sentiments usually have no orientations, and there are also many words whose semantic orientations depend on their contexts. To keep our approach simple, in this paper we only deal with those opinion words that have explicit positive or negative sentiments. However, to capture those opinion words that might not have appeared in the review database, we make use of the adjective synonym set and antonym set in WordNet [42] to generate a larger set of opinion words. In general, adjectives share the same orientation as their synonyms and opposite orientations as their antonyms. In our approach, we first identify all opinion words related to product features and then use them as seed adjectives to search for more adjectives from WordNet. In addition, this set of words is further evaluated by existing tools (e.g., thesaurus.com) to derive additional opinion terms. The synonyms of positive words are amended to the set of positive words, likewise for negative words.

Finally, all collected opinion words are labeled by product features and orientations, which are stored in the opinion words database as shown in Fig. 1.

5.4. *Opinion Orientation Identification*

Once we have defined the major product features along with their product feature sets and opinion word sets, we can start analyzing each product review to determine its opinion orientations on the selected product features. The resulting information about the opinion orientations will be stored in two matrices, the matrix *Positive Feature Count* (*PFC*) and the matrix *Negative Feature Count* (*NFC*), both of dimension $n \times k$, where n is the total number of reviews and k is the total number of selected product features. Note that this feature count information is further processed by considering the reliability of each review when calculating their mass values. When one element from the matrix *PFC* or *NFC* is equal to 1, it indicates that some review supports a particular product feature to be good or bad, respectively. For example, if $PFC_{r,j} = 1$, it means that review r can be considered as evidence supporting that feature j is favorable; otherwise, if $PFC_{r,j} = 0$, it means there is no evidence supporting that feature j is favorable in review r. On the other hand, if $NFC_{r,j} = 1$, it implies that review r can be considered as evidence supporting that feature j is unfavorable; otherwise, if $NFC_{\mathrm{r},\ \mathrm{j}} = 0$, it implies there is no evidence supporting that feature j is unfavorable in review r. Note that when a review is considered as evidence supporting that a particular product feature is favorable or unfavorable, it is also the corresponding evidence supporting that the product is favorable or not. Algorithm 1 is used to compute the feature count matrices *PFC* and *NFC*. In this algorithm, we first initialize all elements in feature count matrices *PFC* and *NFC* to 0. Then, we go through each review r to identify whether it supports that feature j is favorable or unfavorable. If the review supports feature j as favorable, $support_j$ is set to 1; otherwise, it is set to 2. If $support_j$ remains 0 after the feature counting process, it means that there is no evidence in the review that product feature j is favorable or not. If review r supports that product feature j is both favorable and unfavorable, there must exist a contradiction in the review. In this case, we disregard the contradicting opinion orientations on product feature j in review r, and set $support_j$ to 0, accordingly. In addition, if a review supports that a particular product feature is favorable or unfavorable multiple times, we count it as a single piece of evidence; thus, all elements in matrices *PFC* and *NFC* must be either 0 or 1.

To make the algorithm easy to understand, we assume that no negation words or phrases appear in the review sentences. However, when negation words or phrases are present in a sentence, the opinion orientations expressed in the sentences must be reversed; in this case, we must apply the negation rules to correctly count the product features. Examples of negation words or phrases include traditional words such as "*no*," "*not*," and "*never*," as well as pattern-based negations such as "*stop*" + "*vb-ing*," "*quit*" + "*vb-ing*," and "*cease*" + "*to vb*," where *vb* is the POS tag for verb and "*vb-ing*" is *vb* in its

-*ing* form. The following examples show some negation rules that can be used to correctly count product features in a review:

Negation Negative → Positive // e.g., "no problem"
Negation Positive → Negative //e.g., "not good"
Negation Neutral → Negative //e.g., "does not work"

Once we have determined the correct opinion orientations of a review for the selected product features, the corresponding elements in the matrix *PFC* or *NFC* can be updated and stored.

Algorithm 1: Feature Count

Input: A set of *n* reviews *rSet* for product *p* with *k* features
Output: Positive feature count matrix *PFC* and negative feature count matrix *NFC*, both with dimension $n \times k$.

```
Initialize matrices $PFC$ and $NFC$ to 0
for each feature $j$
    Retrieve product feature set $fSet_j$ for feature $j$
    Retrieve positive opinion word set $pSet_j$ and negative opinion word set $nSet_j$ for feature $j$
for each review $r$ in $rSet$
    Detect sentences, and store all sentences in set $sSet$
    Initialize $k$-dimension integer array $support$ to 0
    Initialize $k$-dimension boolean array $contradiction$ to false
    for each sentence $s$ in $sSet$
        for each feature $j$
            if $s$ contains any elements in both $fSet_j$ and $pSet_j$
                then if ($support_j$ == 2) then $contradiction_j$ = true
                    else $support_j$ == 1  // feature $j$ is favorable
            else if $s$ contains any elements in both $fSet_j$ and $nSet_j$
                then if ($support_j$ == 1) then $contradiction_j$ = true
                    else $support_j$ == 2  // feature $j$ is unfavorable
    for each feature $j$
        if $contradiction_j$ == true then $support_j$ = 0
        if $support_j$ == 1 then $PFC_{r,j}$ = 1
        else if $support_j$ == 2 then $NFC_{r,j}$ = 1
Store matrix $PFC$ and $NFC$
```

6. A Cost-Effectiveness Analysis Model

6.1. *Reliability of Product Reviews*

Before calculating the basic mass assignments for product reviews, we must first compute the reliability of each review. The reliability of a review is determined by a number of factors, called *review properties*, which are important indicators to assess whether a review can be trusted or not [8]. We now define the review reliability function as follows.

Definition 3: *Review reliability* is defined as a function *Rel*: $R \rightarrow [0, 1]$, where *R* is a set of reviews, each of which is mapped to a real value between 0 and 1. Each review $r \in R$ is defined as a 3-tuple (*T*, *S*, *P*), where *T* is the text of the product review, *S* is the star rating of *r*, and *P* is a set of review properties defined by a particular e-commerce website. Note that in our approach, only the set of review properties *P* is used to compute the review reliability function.

In the following, we use the Amazon website as an example to illustrate how to calculate review reliability. The Amazon website allows a review to be voted as a *helpful* review by its customers, which is defined as *Helpful Rate* (rp_1) in this paper. The more votes as helpful reviews, the more reliable the review is. Let the maximum number of votes for all reviews be *max_votes* and the number of helpful votes be *helpful_votes*, we calculate the *Helpful Rate* (rp_1) of the review as *help_votes* / *max_votes*. Note that for a typical online product, it is reasonable to assume *max_votes* > 0; otherwise, when *max_votes* = 0, we set $rp_1 = 0.5$.

We further identify four additional review properties as factors that contribute to the calculation of review reliability. Those factors are *Purchased* (rp_2), *Date* (rp_3), *Badges* (rp_4), and *Top Reviewer Ranking* (rp_5). Note that the properties rp_2, rp_4, and rp_5 are properties of the reviewer who wrote the review. Since the reliability of a review is closely related to the reliability of the person who wrote it, in this paper we also consider them properties of the review. We now provide the detailed descriptions of the review properties rp_2 to rp_5 as follows.

Purchased (rp_2) is a label for a reviewer that indicates the e-commerce company has verified the reviewer has purchased the product. A reviewer who has made a purchase and had a real experience with the product can certainly write more reliable reviews than those who have not made a purchase.

Date (rp_3) is the date when the review was posted. For simplicity, we convert the date into the number of months that have passed since the review was posted. The more recent a review was written, the more useful and reliable the review is.

Badges (rp_4) is the number of badges a reviewer has been awarded. At Amazon website, there are a total of 15 types of badges. For example, the REAL NAME badge indicates that the customer used the real name as appeared on the customer's credit card. The more badges a reviewer owns, the better review history the reviewer should have.

Top Reviewer Ranking (rp_5) of a reviewer reflects the opinions of other customers about the reviewer. A reviewer's *Top Reviewer Ranking* is determined by the overall helpfulness of the reviewer's reviews, factoring in the number of reviews the reviewer has written.

Note that the above review properties are the only properties available on Amazon website that relate to review reliability. Before calculating the reliability of a review, we need to normalize the property values to values in range [0, 1]. Table 1 shows the value ranges and the normalized values for the five review properties rp_1 to rp_5.

Table 1. A List of Five Review Properties Used to Determine the Review Reliability.

Property	Description	Range	Normalized Value
rp_1	Helpful Rate	[0, 1]	helpful_votes / max_votes 0.5 (if max_votes = 0)
rp_2	Purchased	{0, 1}	0 → 0 (not purchased) 1 → 1.0 (purchased)
rp_3	Date	[0, +∞)	0~3 months → 1.0 3~6 months → 0.7 6~12 months → 0.4 > 1 year → 0.1
rp_4	Badges	[0, 15]	num_of_badges / 15
rp_5	Top Reviewer Ranking	[1, +∞)	< 1,000 → 1.0 1,000~10,000 → 0.7 >10,000 → 0.5

In Table 2, we give a few examples of raw data collected from the Amazon website, where each row contains five review properties associated with a product review, as well as the person who wrote the review. The normalized values of the review properties are shown in parentheses following the raw data, and these values are computed according to the conversion rules defined in Table 1.

Table 2. Examples of Collected Raw Review Property Data.

Review ID	Helpful Votes / Max Votes	Purchased	Date	Badges	Top Ranking
1	118/118 (1.0)	1 (1)	17 (0.1)	1 (0.53)	63,027 (0.5)
2	83/118 (0.703)	1 (1)	5 (0.7)	2 (0.56)	556 (1.0)
3	71/118 (0.602)	0 (0)	16 (0.1)	0 (0.50)	81,258 (0.5)
4	26/118 (0.5)	1 (1)	11 (0.4)	3 (0.60)	292,053 (0.5)
5	98/118 (0.831)	1 (1)	16 (0.1)	0 (0.50)	458,571 (0.5)

Note that in order to normalize the findings, we introduced a base value of 0.5 for the badges, helpful rate and top reviewer ranking properties. Thus, for these properties, the lowest value is 0.5 and only higher values are allowed. The reliability *Rel*(*r*) of a product review *r* can be calculated as in Eq. (6).

$$Rel(r) = w_1 \times (r.rp_1 + r.rp_2 + r.rp_3 + r.rp_4 + r.rp_5) \tag{6}$$

where the weight w_1 indicates the importance of the review properties rp_1 to rp_5. Existing work on interpreting the significance of online reviews at Amazon website has used these properties in a similar way. For example, Li *et al.* discussed about the relation between review helpfulness and spam review for identification of review spam [21], and Mukherjee *et al.* considered AVP (Amazon Verified Purchase) property for spotting fake reviews [23]. Both spam and fake reviews are unreliable reviews in evaluating product quality. In our approach, we selected five major review properties, including review helpfulness (rp_1) and verified purchase (rp_2), to measure the reliability of product reviews. Since there is no clear evidence showing any review property would be more important than any others in evaluating review reliability, we consider each review property to have an equal weight and set w_1 to 0.2. Due to the normalization of each review property value, *Rel*(*r*) calculated using Eq. (6) will be a reliability value in the range of [0, 1].

6.2. *Calculation of Basic Mass Assignments*

To calculate the BMAs of both positive and negative counts for feature *j* of product $p \in P$, we compute the weighted average *PFC* elements (WAP_j) and weighted average *NFC* elements (WAN_j) for the positive and negative orientations, respectively, as in Eqs. (7) and (8).

$$WAP_j = \frac{PFC_{r_1,j} \times Rel(r_1) + PFC_{r_2,j} \times Rel(r_2) + \ldots + PFC_{r_n,j} \times Rel(r_n)}{\sum_{i=1}^{n} PFC_{r_i,j}} \tag{7}$$

$$WAN_j = \frac{NFC_{r_1,j} \times Rel(r_1) + NFC_{r_2,j} \times Rel(r_2) + \ldots + NFC_{r_n,j} \times Rel(r_n)}{\sum_{i=1}^{n} NFC_{r_i,j}} \tag{8}$$

where $Rel(r_i)$ is the review reliability of review r_i calculated according to Eq. (6), $1 \le i \le n$, and *n* is the total number of reviews for product *p*. For a given product feature *j*, both WAP_j and WAN_j must fall in the range [0, 1] since the reliability of each review is in the range [0, 1], and the maximum count of each review is 1 in either positive or negative orientation.

Let *F* = {*favorable*} and ~*F* = {*unfavorable*}, we have $U = F \cup {\sim}F$ = {*favorable*, *unfavorable*}. For product feature *j*, the BMAs of the positive feature counts and negative feature counts can be calculated as in Eqs. (9) and (10). Note that $m_{POS}(U)$ and $m_{NEG}(U)$ refer to the mass values of uncertainty for positive feature counts and negative feature counts, respectively.

$$\begin{cases} m_{POS\_j}(F) = WAP_j \\ m_{POS\_j}(\sim F) = 0 \\ m_{POS\_j}(U) = 1 - WAP_j \end{cases} \quad (9) \qquad \begin{cases} m_{NEG\_j}(F) = 0 \\ m_{NEG\_j}(\sim F) = WAN_j \\ m_{NEG\_j}(U) = 1 - WAN_j \end{cases} \quad (10)$$

Since the number of reviews can be a good indicator of a product's popularity, we consider it as special independent evidence supporting whether the product is favorable or not. To assess how the number of reviews has an impact on product favorability, we first identify the maximum number of reviews (denoted as NR_{max}) among all product alternatives in *P*. Then we compare the number of reviews of product *p* with NR_{max} to quantify its impact on the belief value of product *p*. We realize that some popular product may have an extremely large number of reviews when compared to others. In this case, the result will be dominated by this value, leaving less popular but potentially better-quality products behind. To mitigate this domination, we use a logarithm function to scale our mass assignments and narrow the gaps between the numbers of reviews for different product alternatives. We use the following simple example to illustrate the basic idea. Suppose in a set of product alternatives P, NR_{max} = 1,999, and for a particular product $p \in P$, the number of reviews NR_p is equal to 99. When we compare NR_p with NR_{max}, the impact of NR_p becomes very small, even though 99 positive reviews represent a considerable number of reviews. Now if we try to compare lgNR_p with lgNR_{max}, the gap between them can be significantly reduced and the number of reviews, NR_p = 99, can be duly considered as evidence supporting that p is a favorable product.

The BMAs for the number of reviews of product p can be calculated as in Eqs. (11-13), where $NR_{max} > 0$.

$$m_{NR}(F) = \begin{cases} \dfrac{2 \times \lg(NR_p + 1)}{\lg(NR_{\max} + 1)} - 1 & if\ \lg(NR_p + 1) \geq (\lg(NR_{\max} + 1))/2 \\ 0 \quad otherwise & \end{cases} \tag{11}$$

$$m_{NR}(\sim F) = \begin{cases} 1 - \dfrac{2 \times \lg(NR_p + 1)}{\lg(NR_{\max} + 1)} & if\ \lg(NR_p + 1) < (\lg(NR_{\max} + 1))/2 \\ 0 \quad otherwise & \end{cases} \tag{12}$$

$$m_{NR}(U) = 1 - m_{NR}(F) - m_{NR}(\sim F) \tag{13}$$

Note that to deal with the special cases when NR_p = 0 and/or NR_{max}= 1, in Eqs. (11) and (12), we replace NR_p and NR_{max} with (NR_p+1) and (NR_{max}+1), respectively. When lg(NR_p +1) ≥ (lg(NR_{max}+1))/2, we consider it as evidence supporting that the product is favorable. As special cases, when NR_p = NR_{max}, the mass value $m_{NR}(F)$ is equal to 1, which means the evidence fully supports that the product is favorable. When $\log_{10}$(NR_p +1) = ($\log_{10}$(NR_{max}+1))/2, the mass value $m_{NR}(F)$ is equal to 0, which means we do not consider the insufficient number of reviews as evidence supporting that the product is favorable. On the other hand, when lg(NR_p+1) < (lg(NR_{max}+1))/2, we consider it as evidence supporting that the product is unfavorable rather than favorable. This is because few product reviews usually indicate that the product is not popular. As a special case, when NR_p = 0, the mass value $m_{NR}(\sim F)$ is equal to 1. Since no one has given any comment to the product yet, we prefer not to recommend it to customers for purchasing.

6.3. *Combination of Evidence*

Once the BMAs for all the evidence are calculated, they can be combined in a systematic way to provide a more complete assessment of product effectiveness by reducing the uncertainty associated with individual pieces of evidence. The evidence fusion procedure uses Dempster's rule of combination. As shown in Fig. 1, we first combine each feature's positive counts and negative counts into masses m_j for product feature j, where $1 \leq j \leq k$, and k is the total number of selected product features. The corresponding evidence combination rules for F, $\sim F$ and U are listed as in Eqs. (14-16).

$$m_j(F) = m_{POS\_j}(F) \oplus m_{NEG\_j}(F) \tag{14}$$

$$m_j(\sim F) = m_{POS\_j}(\sim F) \oplus m_{NEG\_j}(\sim F) \tag{15}$$

$$m_j(U) = m_{POS\_j}(U) \oplus m_{NEG\_j}(U) \tag{16}$$

When the masses for all product features and the number of reviews are calculated, we can again use Dempster's rule of combination to combine them into masses m_p for product p as in Eqs. (17-19). It is important to note that the evidence of product features is combined one by one to derive their joint masses, and finally combined with the evidence of the number of reviews to derive the masses for the product.

$$m_p(F) = m_1(F) \oplus m_2(F) \oplus \ldots \oplus m_k(F) \oplus m_{NR}(F) \tag{17}$$

$$m_p(\sim F) = m_1(\sim F) \oplus m_2(\sim F) \oplus \ldots \oplus m_k(\sim F) \oplus m_{NR}(\sim F) \tag{18}$$

$$m_p(U) = m_1(U) \oplus m_2(U) \oplus \ldots \oplus m_k(U) \oplus m_{NR}(U) \tag{19}$$

According to Eq. (3), the belief values for the product hypotheses can be calculated as in Eqs. (20) and (21).

$$belief_p(F) = m_p(F) \tag{20}$$

$$belief_p(\sim F) = m_p(\sim F) \tag{21}$$

We now use an example to show how the joint masses for combined evidence can be calculated. Suppose we want to calculate the mass values for the product feature "Sound" (denoted as S). According to Eq. (5), we can calculate $m_S(F)$, $m_S(\sim F)$ and $m_S(U)$ as in Eqs. (22-24).

$$m_S(F) = m_{POS\_S}(F) \oplus m_{NEG\_S}(F) = \tag{22}$$

$$\frac{m_{POS\_S}(F) \times m_{NEG\_S}(F) + m_{POS\_S}(F) \times m_{NEG\_S}(U) + m_{POS\_S}(U) \times m_{NEG\_S}(F)}{1 - C}$$

$$m_S(\sim F) = m_{POS\_S}(\sim F) \oplus m_{NEG\_S}(\sim F) = \tag{23}$$

$$\frac{m_{POS\_S}(\sim F) \times m_{NEG\_S}(\sim F) + m_{POS\_S}(\sim F) \times m_{NEG\_S}(U) + m_{POS\_S}(U) \times m_{NEG\_S}(\sim F)}{1 - C}$$

$$m_S(U) = m_{POS\_S}(U) \oplus m_{NEG\_S}(U) = \frac{m_{POS\_S}(U) \times m_{NEG\_S}(U)}{1 - C} \tag{24}$$

where $C = m_{POS\ S}(F) \times m_{NEG\ S}(\sim F) + m_{POS\ S}(\sim F) \times m_{NEG\ S}(F)$. Note that since $U \cap F = F$ and $U \cap \sim F = \sim F$, we have $U \cap F \neq \emptyset$ and $U \cap \sim F \neq \emptyset$ in Eqs. (22-23). The joint mass values for the other product features and the joint mass values for product *p*, i.e., $m_p(F)$, $m_p(\sim F)$ and $m_p(U)$, can be calculated in a similar way.

According to Eqs. (20) and (21), the belief value indicating that a product *p* is favorable is equal to $m_p(F)$, and the belief value indicating that a product *p* is unfavorable is equal to $m_p(\sim F)$. The uncertainty that product *p* is both favorable and unfavorable can be quantified by (1 - ($belief_p(F)$ + $belief_p(\sim F)$)). By considering uncertainty, we can calculate the effectiveness of product *p*, denoted as *E*(*p*), by summing its belief value for being favorable and an adjustment Δ_p equal to 50% of the uncertainty value, as in Eqs. (25) and (26).

$$\Delta_p = 0.5 \times (1 - (belief_p(F) + belief_p(\sim F))) = 0.5 \times m_p(U) \tag{25}$$

$$E(p) = Bel(p) = belief_p(F) + \Delta_p = m_p(F) + 0.5 \times m_p(U) \tag{26}$$

By further taking the price factor into consideration, we can calculate the cost-effectiveness value of product *p*, denoted as *CE*(*p*), as in Eq. (27).

$$CE(p) = E(p) / Cost(p) \tag{27}$$

where *Cost*(*p*) is the normalized cost of product *p*. Let $P = \{p_1, p_2, \ldots, p_n\}$ be a set of product alternatives to be evaluated and ranked, which have similar functionality and are in the same price range. For $\forall\ p \in P$, *Cost*(*p*) can be calculated as in Eq. (28).

$$Cost(p) = mCost(p) / Max(mCost(p_1), mCost(p_2), ..., \ mCost(p_n)) \tag{28}$$

where *mCost*(*p*) is the cost function (defined in Section 4.1) that maps product *p* to the lowest price offered by one of the online sellers. With the *CE* value of each product in set *P*, we can rank the product alternatives and provide users useful insights about the quality and popularity of the products being shopped online.

7. Case Studies

In this section, we demonstrate how our analytical model based on D-S theory can be used to analyze data sets collected from Amazon website. We use two case studies to illustrate how our analytical model can provide more reliable and accurate results than the commonly used ASR-based product ranking mechanism. The data used in the case studies were collected from Amazon’s product records, but note that the product data, such as star ratings, minimal price, and the number of reviews may have changed by the time of this publication.

7.1. *Case Study 1: Digital Camera*

In this case study, we collected 10 digital camera products priced between $300-$700, all of which have different brands, series, star ratings and number of reviews, but whose ASRs are at least 4.0. To assess the quality of products in the category of "Digital Camera," we selected 6 major product features, namely "Picture," "Battery Life," "Zoom," "Video," "Lens," and "Shutter Speed." Note that potential product features were detected through frequency analysis of review text and refined using the domain knowledge such as professional listings and external review sites such as Consumer Reports. The features that were well represented in the reviews of most items were then selected for use as the main product features. We analyzed the product reviews of each product using our text mining approach and counted the positive and negative orientations for each product feature. Table 3 shows the feature count information for 10 digital camera products, where the brand of each camera product and some additional product information can be found in Table 4.

Table 3. Feature Count Information of the Ten Digital Camera Products.

Item #	Picture		Battery Life		Zoom		Video		Lens		Shutter Speed	
	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
1	5	0	1	1	0	0	6	0	4	0	1	0
2	58	2	4	2	11	2	31	0	43	3	13	0
3	72	3	11	6	35	8	54	3	33	10	10	4
4	190	4	20	1	10	1	22	2	52	3	10	2
5	309	5	28	1	21	2	40	2	109	3	16	1
6	16	0	2	0	3	0	1	0	2	0	0	1
7	61	5	4	1	41	1	13	1	9	2	3	1
8	58	5	8	1	1	2	39	8	12	3	6	1
9	148	24	49	10	13	4	41	4	9	1	2	2
10	60	2	9	1	37	3	6	1	16	1	5	1

In Table 4, we list these 10 digital camera products, along with some additional product information and analysis results. For each product, "ASR" refers to the average star rating of the product posted on its corresponding product page; "# of Reviews" is the total number of reviews including both positive (4 and 5 stars) and negative (1, 2, and 3 stars) reviews; and "Price" refers to the price immediately available, either for Amazon Prime or for individuals, offering the product as new. The last two columns show the *E* values and *CE* values generated for each item by the analytical model presented in the previous section, where an *E* value quantifies the quality and popularity of the product, and a *CE* value quantifies the cost-effectiveness. Based on the ASR ranking, the three product alternatives, No.10, No. 4 and No. 5, have the highest ASR values of 4.6, 4.7 and 4.8, respectively, and appear to be the best choices for purchase. Among these 3 top products, it may be difficult for customers to choose which one to buy because they all have similar ASR values and a considerable number of product reviews. If the customer prefers the most popular one, and is not too concerned about the price, the customer is likely to choose product No. 5 for purchase.

Table 4. Product Information of Ten Digital Cameras and the Analysis Results.

Item #	ASR	# of Reviews	Price	Product & Brand	E-Value	CE-Value
1	4.4	17	349.45	TomTom Bandit	**0.758**	**1.51**
2	4.3	138	697.99	Panasonic LUMIX DMC-LX100K	0.729	0.729
3	4.3	214	697.99	Panasonic LUMIX DMC-FZ1000	0.276	0.276
4	**4.7**	612	399	Canon EOS Rebel T5	0.466	0.815
5	**4.8**	854	546.95	Nikon D3300	**0.873**	**1.11**
6	4.1	41	329	Fujifilm FinePix S9900W	**0.953**	**2.02**
7	4.3	190	339.95	Nikon COOLPIX S9900	0.313	0.643
8	4.0	195	346.95	Ricoh Theta S Digital Camera (Black)	0.312	0.628
9	4.0	416	349.95	Canon Powershot A1200	0.5	0.997
10	**4.6**	146	422.70	Nikon COOLPIX P610	0.279	0.461

Now with our analytical model, we can look into the *E* values and the *CE* values for all 10 product alternatives. Since the *E* value quantifies the quality level as well as the popularity of the product, a customer who only cares about quality and popularity may choose the products with high effectiveness values. The top three choices are No. 6 with *E* value 0.953, No. 5 with *E* value 0.873, and No. 1 with *E* value 0.758. On the other hand, if the customer is concerned about both effectiveness and cost, the customer may choose the products with high *CE* values. In this case, the top three choices are No. 6 with *CE* value 2.02, No. 1 with *CE* value 1.51, and No. 5 with *CE* value 1.11. Note that the ranking results calculated using our analytical model differ from the ASR-based ranking results; however, our ranking results are more *accurate* and *reliable* for online shopping because our model takes into account sufficient evidential information before the ranking results are calculated.

Since there is no ground truth about the real quality of online products, to validate our analysis results, we examined the raw data collected for our case study. We noticed that although product No. 4 has a fairly high ASR, its effectiveness value is merely 0.466. By investigating its raw review comments, we found that a considerable number of reviews were written by unreliable reviewers who wrote reviews with very low helpfulness rates. As can be seen from this example, while ASR is vulnerable to unhelpful and possibly even malicious reviewers, our effectiveness metrics place less weight on these opinions and more on the opinions of those who are reliable and established reviewers. Moreover, by investigating the feature counts in Table 3, we saw that product No. 4 had more negative opinion counts than product No. 6. The above situation leads to a lower effectiveness value for product No. 4 compared to product No.

6. Product No. 5 and No. 6 have very dissimilar ASR values, 4.8 (the highest ASR value) and 4.1 (the second lowest), respectively. However, the *E* value of product No. 6 is greater than that of product No. 5, at 0.953 and 0.873, respectively, even though No. 6 has fewer reviews, indicating that it is not as popular as No. 5. By examining the feature counts for both products in Table 3, we found that product No. 5 had negative reviews for all product features. No. 6, on the other hand, had only one negative feature count. Thus, as determined by feature-based evidence, product No. 6 has higher quality than product No. 5, and product No. 6's overall effectiveness value and low price make it the best choice for online purchase.

7.2. *Case Study 2: Audio/Video (A/V) Receiver*

For the second case study, we collected 10 A/V receiver products in the price range of $300-$600, all with different brands, series, star ratings and number of reviews, but they all had ASRs of at least 3.8. To assess the quality of products in the category of "A/V Receiver," we selected 6 major product features, namely "Sound," "HDMI," "Setup," "Remote Control," "Warranty," and "Digital Connection." We analyzed the product reviews of each using our text mining approach and counted the positive orientations and negative orientations for each product feature. Table 5 shows the feature count information for the 10 A/V receiver products, where the brand of each receiver product as well as some additional product information can be found in Table 6. Based on the ASR ranking in Table 6, the top three product alternatives are No. 6 (with ASR 4.5), No. 7, No. 9 or No. 10 (each with ASR 4.3) and No. 1, No. 3 or No. 5 (each with ASR 4.2). These seem to be the best choices for online purchases. By further examining the number of reviews and the lowest prices, a customer may choose one of the seven options (e.g., a customer might choose product No. 9 for a purchase because of its high ranking, reasonable price and high number of reviews).

Table 5. Feature Count Information of the Ten A/V Receivers.

Item #	Sound		HDMI		Setup		Remote Control		Warranty		Digital Connect	
	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
1	121	3	174	15	20	2	85	3	2	2	59	1
2	33	3	40	8	9	1	19	4	2	0	6	0
3	45	2	54	6	11	1	28	3	1	0	14	1
4	30	3	31	5	4	0	22	3	0	0	8	1
5	13	0	20	1	4	0	11	0	0	0	4	0
6	5	0	1	0	1	1	3	0	0	0	0	0
7	13	0	11	1	1	0	11	2	0	1	3	0
8	170	6	196	24	25	9	79	8	1	0	12	0
9	221	5	291	11	39	8	110	2	2	1	85	2
10	71	0	63	5	11	0	32	0	0	1	25	2

Table 6. Product Information of Ten A/V Receivers and the Analysis Results.

Item#	ASR	# of Reviews	Price	Product & Brand	E-Value	CE-Value
1	**4.2**	484	369.95	Yamaha RX-V677	0.502	0.813
2	4.1	115	569.88	Onkyo TX-NR727	0.612	0.643
3	**4.2**	156	549.95	Yamaha RX-A840BL	0.44	0.479
4	4.1	97	379	Denon AVR-S710W	0.583	**0.922**
5	**4.2**	45	599	Denon AVR-X1200W	**0.898**	0.898
6	**4.5**	7	370	Sony STR-DA3200ES ES	0.483	0.782
7	**4.3**	36	349.95	Onkyo TX-NR545	0.524	0.897
8	3.8	603	510.57	Onkyo TX-SR313	**0.882**	**1.040**
9	**4.3**	728	498.95	Yamaha RX-V675	**0.766**	0.920
10	**4.3**	188	329.95	Yamaha RX-V477	0.677	**1.230**

Now with our analytical model, we can calculate the *E* values and *CE* values for all 10 products, which are listed in the last two columns of Table 6. Since the effectiveness value quantifies the quality level and the popularity of the products, a customer who only cares about quality and popularity may choose products with high effectiveness values. The top three choices are No. 5 with an *E* value of 0.898, No. 8 with an *E* value of 0.882, and No. 9 with an *E* value of 0.766. On the other hand, if the customer is concerned about both effectiveness and cost, the customer may choose products with high *CE* values. In this case, the top three choices are No. 10 with an *CE* value of 1.230, No. 8 with an *CE* value of 1.040, and No. 4 with an *CE* value of 0.922. Note that our evidence-based approach produces ranking results that are quite different from those produced by traditional ASR-based ranking mechanisms.

To validate the results of our analysis, we examined the raw data collected for our case study. For product No. 6, although it has the highest ASR (4.5), it does not have a high enough effectiveness value compared to other product alternatives because it has only 7 product reviews (one of which is negative for "Setup"). Products No. 7, No. 9 and No. 10 have very good ASRs and No. 9 has the highest number of reviews. However, No. 7 has negative reviews for several features and a low number of reviews. While the ASR is not sufficient to distinguish No. 7 from other highly rated products, our metrics accounting for sentiment and review reliability indicate that No. 7 has the lowest *E* value of the three. As shown in Table 5 and Table 6, product No. 10 has a higher number of reviews (although significantly less than the highest) and these are largely positive in orientations. However, when examining its review properties, we found that many of the reviews were not reliable (e.g., the majority of positive reviews have no *Helpful* votes). This result demonstrates the strength of our approach in going beyond simple sentiment analysis to incorporate the reliability of reviews. In addition, from Table 5, product No. 10 has a feature, "Warranty", that contains a negative feature count but zero positive feature count; in general, its effectiveness value is significantly affected, resulting in a low *E* value (0.677) for this product. On the other hand, product No. 5 has a relatively low ASR. It also has the third lowest number of reviews. However, due to only one negative feature review and high *Helpful Rates*, product No. 5 has the highest *E* value among the 10 product alternatives. Thus, if a customer is not too concerned about price, the customer should choose product No. 5 for purchase. However, if price is a major issue to consider, product No. 8 with a good *E* value (0.882), can be chosen, even though its *CE* value is not the highest.

8. Conclusions and Future Work

In this paper, we introduce an evidence-based approach to evaluating and ranking online products in e-commerce. By calculating the product effectiveness and the cost-effectiveness values for various product alternatives sold online, we have developed a formal cost-effectiveness analysis model using D-S theory. In our approach, we use the reviews of products and their review properties as evidence to justify whether a product is a favorable one or not. Product data from e-commerce websites such as Amazon is quantified and evaluated using our formal methodology. By applying Dempster's rule of combination, we combine different pieces of evidence to derive more reliable belief values about the effectiveness of the products. As such, our analytical model can properly handle uncertain information, reduce the degree of uncertainty, and produce more reliable and accurate results than conventional ranking mechanisms, such as those based on ASR. Our case studies show that by ranking the product alternatives based on effectiveness and cost-effectiveness values, our approach can be very helpful in assisting customers make the right purchase decisions.

In future research, we plan to consider more factors, including review lengths, and apply data mining methods to classify product reviews into more meaningful groups. To improve accuracy and handle subtle situations, such as negative comments using positive words, we will consider using word embeddings and deep neural networks to analyze review comments. Existing approaches, such as BERT [31], can complement our evidence-based approach and be used to improve the initial sentiment analysis. In addition, ablation studies can be performed to show the impact of each component of the work on the performance of the proposed approach. As the classified evidence is used as independent evidence for evidence combination, it may further help to reduce the level of uncertainty and lead to more accurate and reliable product ranking results. We will also investigate other domains, such as medical and healthcare services and mobile apps, to explore the opportunities of evaluating online services and apps using our proposed evidence-based approach. Finally, we will address the limitations associated with manual determination of product features in future research. With automated product feature extraction, we will be able to apply our approach to large sets of products to properly evaluate the efficiency of our proposed method and further illustrate how *E* value can accurately quantify the quality level of a product as well as its popularity. As some preliminary efforts along this direction, we have started to experiment with deep neural networks to automatically extract product features from online product reviews [43].

Acknowledgements

We thank the editors and all anonymous referees for the careful review of this paper and the many suggestions for improvements they provided. This material is based upon work partially supported by the President's 2017 Creative Economy (CE) Fund, University of Massachusetts.

References

[1] Amazon, *Submit a Review*, Amazon.com Help, 2014. Retrieved on October 7, 2021 from https://www.amazon.com/gp/help/customer/display.html?nodeId=GL4WJF8BGV8VL6B8

[2] P. A. Pavlou and A. Dimoka, "The nature and role of feedback text comments in online marketplaces: implications for trust building, price premiums, and seller differentiation," *Information Systems Research*, Vol. 17, No. 4, 2006, pp. 392-414.

[3] Q. Ye, R. Law, and B. Gu, "The impact of online user reviews on hotel room sales," *International Journal of Hospitality Management,* Vol. 28, No.1, 2009, pp. 180-182.

[4] M. McGlohon, N. Glance, and Z. Reiter, "Star quality: aggregating reviews to rank products and merchants," *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM10)*, Washington DC, May 2010, pp. 114-121.

[5] N. Hu, P. A. Pavlou, and J. Zhang, "Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of online word-of-mouth communication," *Proceedings of the 7th ACM Conference on Electronic Commerce (EC' 06)*, Ann Arbor, MI, USA, June 11 - 15, 2006, pp. 324-330.

[6] G. Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, NJ, USA, 1976.

[7] J. Kohlas and P.A. Monney, *A Mathematical Theory of Hints: An Approach to the Dempster-Shafer Theory of Evidence*, Lecture Notes in Economics and Mathematical System, Springer-Verlag Berlin Heidelberg New York, 1995.

[8] R. Wei and H. Xu, "A formal cost-effectiveness analysis model for product evaluation in e-commerce," *Proceedings of the 25th International Conference on Software Engineering and Knowledge Engineering (SEKE 2013),* Boston, MA, USA, June 27-29, 2013, pp. 287-293.

[9] P. Walley, *Statistical Reasoning with Imprecise Probabilities*, Chapman & Hall, New York, NY, USA, 1991.

[10] L. A. Zadeh, "A fuzzy-algorithmic approach to the definition of complex or imprecise concepts," *International Journal of Man-Machine Studies*, Vol. 8, No. 3, 1976, pp. 249-291.

[11] M. Richardson and P. Domingos, "Markov logic networks," *Machine Learning*, Vol. 62, 2006, pp. 107-136.

[12] F. Dong, S. M. Shatz, and H. Xu, "Reasoning under uncertainty for shill detection in online auctions using Dempster-Shafer theory," *International Journal of Software Engineering and Knowledge Engineering (IJSEKE)*, Vol. 20, No. 7, Nov. 2010, pp. 943-973.

[13] S. Panigrahi, A. Kundu, S. Sural, and A. K. Majunmdar, "Use of Dempster-Shafer theory and Bayesian inferencing for fraud detection in mobile communication networks," *Lecture Notes in Computer Science*, Spring Berlin Heidelberg, Vol. 4586, 2007, pp. 446-460.

[14] R. Zhang, K. Wang, and C. Chen, "Service supplier selection decision based on Dempster–Shafer synthesis rule," *Journal of Industrial Integration and Management*, Vol. 4, No. 2, 2019, 1950004.

[15] P. Li and C. Wei, "An emergency decision-making method based on D-S evidence theory for probabilistic linguistic term sets," *International Journal of Disaster Risk Reduction*, Vol. 37, 2019, 101178.

[16] J. Yang and G. Mandan, "An evidential reasoning approach for multiple attribute decision making with uncertainty," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 24, No. 1, Jan. 1994, pp. 1-17.

[17] Y. H. Cho and J. K. Kim, "Application of web usage mining and product taxonomy to collaborative recommendations in e-commerce," *Expert Systems with Applications*, Vol. 26, No. 3, 2004, pp. 234-246.

[18] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Analysis of recommendation algorithms for e-commerce," *Proceedings of the 2nd ACM Conference on Electronic Commerce (EC'00)*, Minneapolis, MN, USA, October 17-20, 2000, pp.158-167.

[19] Z. Zhang and J. Feng, "Price of identical product with gray market sales: an analytical model and empirical analysis," *Information Systems Research*, Vol. 28, No. 2, April 2017, pp. 397-412.

[20] G. Wang, S. Xie, B. Liu, and P. S. Yu, "Review graph based online store review spammer detection," *Proceedings of the IEEE 11th International Conference on Data Mining (ICDM-2011)*, Vancouver, Canada, December 11-14, 2011, pp. 1242-1247.

[21] F. Li, M. Huang, Y. Yang, and X. Zhu, "Learning to identify review spam," *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI'11)*, Barcelona, Catalonia, Spain, July 16-22, 2011, pp. 2488-2493.

[22] Z. Wang, S. Gu, and X. Xu, "GSLDA: LDA-based group spamming detection in product reviews," *Applied Intelligence*, Vol. 48, 2018, pp. 3094-3107.

[23] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," *Proceedings of the International World Wide Web Conference (WWW'12)*, Lyon, France, April 16-20, 2012, pp. 191-200.

[24] Y. Heng, Z. Gao, Y. Jiang, and X. Chen, "Exploring hidden factors behind online food shopping from Amazon reviews: A topic mining approach," *Journal of Retailing and Consumer Services*, Vol. 42, 2018, pp. 161-168.

[25] D. Zhu, T. Lappas, and J. Zhang, "Unsupervised tip-mining from customer reviews," *Decision Support Systems*, Vol. 107, March 2018, pp. 116-124.

[26] R. Y. Lau, C. Li, and S. S. Y. Liao, "Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis," *Decision Support Systems*, Vol. 65, Sept. 2014, pp. 80-94.

[27] A. Castillo, D.V. Meer, and A. Castellanos, "ExUP recommendations: inferring user's product metadata preferences from single-criterion rating systems," *Decision Support Systems*, Vol. 108, April 2018, pp. 69-78.

[28] Z. Luo, S. Huang, and K. Q. Zhu, "Knowledge empowered prominent aspect extraction from product reviews," *Information Processing & Management*, Vol. 56, No. 3, 2019, pp. 408-423.

[29] T. U. Haque, N. N. Saber, and F. M. Shah, "Sentiment analysis on large scale Amazon product reviews," *Proceedings of the IEEE International Conference on Innovative Research and Development (ICIRD)*, Bangkok, Thailand, May 11-12, 2018, pp. 1-6.

[30] N. Shrestha and F. Nasoz, "Deep learning sentiment analysis of Amazon.com reviews and ratings," *International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI)*, Vol.8, No.1, February 2019, pp. 1-15.

[31] H. Xu, B. Liu, L. Shu, and P. S. Yu, "BERT post-training for review reading comprehension and aspect-based sentiment analysis," *Proceedings of the 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, Vol. 1, Minneapolis, MN, USA, June 2-7, 2019, pp. 2324-2335.

[32] K. Zhang, R. Narayanan, and A. Choudhary, “Voice of the customers: mining online customer reviews for product feature-based ranking,” *Proceedings of the 3rd Workshop on Online Social Networks (WOSN)*, Boston, MA, USA, June 22, 2010, pp. 11-19.

[33] A. Ghose and P. G. Ipeirotis, “Designing novel review ranking systems: predicting the usefulness and impact of reviews,” *Proceedings of the Ninth International Conference on Electronic Commerce,* August 19-22, 2007, Minneapolis, MN, pp. 303-310.

[34] Z. J. Zha, J. Yu, J. Tang, M. Wang, and T. S. Chua, “Product aspect ranking and its applications,” *IEEE Transactions on Knowledge and Data Engineering,* Vol. 26, No. 5, 2014, pp. 1211-1224.

[35] H. Xu, Y. Zhang, and R. DeGroof, “A feature-based sentence model for evaluation of similar online products,” *Journal of Electronic Commerce Research (JECR)*, Vol. 19, No. 4, November 2018. pp. 320-335.

[36] A. Jøsang and P. Simon, “Dempster’s rule as seen by little colored balls,” *Computational Intelligence,* Vol. 28, No. 4, 2012, pp. 453-474.

[37] A. Jøsang, “The consensus operator for combining beliefs,” *Artificial Intelligence*, Vol. 142, No. 1-2, 2002, pp. 157-170.

[38] F. Campos and S. Cavalcante, “An extended approach for Dempster-Shafer theory,” *Proceedings of the IEEE International Conference on Information Reuse and Integration (IRI 2003)*, Las Vegas, NV, USA, October 27-29, 2003, pp. 338-344.

[39] D. Dubois and H. Prade, “Representation and combination of uncertainty with belief functions and possibility measures,” *Computational Intelligence*, Vol. 4, No. 3, 1988, pp. 244-264.

[40] M. Hu and B. Liu, “Mining and summarizing customer reviews,” *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004),* Seattle, Washington, USA, Aug 22-25, 2004, pp. 168-177.

[41] Apache, *Welcome to Apache OpenNLP*, The Apache Software Foundation, 2017. Retrieved on July 18, 2019 from http://opennlp.apache.org/

[42] E. M. Vorhees, “Using WordNet for text retrieval,” In: C. Fellbaum (ed.), *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA, USA, 1998.

[43] V. Erram, “Automated product feature extraction using recurrent neural networks,” *Master’s Thesis*, Computer and Information Science Department, University of Massachusetts Dartmouth, September 2017.