IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 28, 2020, accepted April 14, 2020, date of publication April 20, 2020, date of current version May 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2988872

Predicting the Helpfulness Score of Product Reviews Using an Evidential Score Fusion Method

FATEMEH FOULADFAR[1], **MOHAMMAD NADERI DEHKORDI**[1], **AND MOHAMMAD EHSAN BASIRI**[2], **(Member, IEEE)**
[1]Faculty of Computer Engineering, Najafabad Branch, Islamic Azad University, Najafabad 1477893855, Iran
[2]Department of Computer Engineering, Shahrekord University, Shahrekord 8818634141, Iran

Corresponding author: Mohammad Naderi Dehkordi (naderi@iaun.ac.ir)

ABSTRACT Everyday many online product sales websites and specialized reviewing forums publish a massive volume of human-generated product reviews. People use these reviews as valuable free source of knowledge when decide to buy products. Therefore, an accurate automated system for distinguishing useful reviews from non-useful ones is of great importance. This article presents a new model for specifying the usefulness of comments using the textual features extracted from the reviews. Various types of features including emotion-related, linguistic and text-related features, valence, arousal, and dominance (VAD) values, review-length and polarity of comments are exploited in this study. Moreover, two new algorithms are presented: an improved evidential algorithm for emotion recognition, and an algorithm for extracting VAD values for each review. Finally, the usefulness of reviews is predicted using the mentioned features and an improved Dempster–Shafer score fusion algorithm. The proposed method is applied to review datasets of Books and Video Games of Amazon. The results show that combining the features associated with emotions, features of VAD, and text-related features improves the accuracy of predicting the usefulness of reviews. Also, in comparison with the original Dempster–Shafer method, the precision of the improved Dempster–Shafer algorithm for both datasets is 15% and 11% higher, respectively.

INDEX TERMS Dempster–Shafer theory, emotion recognition, opinion mining, review helpfulness.

I. INTRODUCTION

With the advent of the Web and the expansion of e-commerce, users are expressing their views on products and services on many specialized and commercial sites to interact and work together. Through online reviews, customers share their personal beliefs, experiences of purchasing decisions, and evaluations towards services or products [1]. These reviews contain valuable information and can be used to analyze people's attitudes and interests. Moreover, they can be used to identify and analyze people's positive and negative views on a variety of targets such as locations, products, and specific events. [2]. Such informative reviews are valuable for both consumers and producers. Consumers read product reviews before making a purchase decision to reduce search costs and purchase uncertainty [3]. A well-written review identifies the strengths and weaknesses of products for producers and identifies what can be learned about new product development [1].

User reviews are already taking up a lot of space on the web and the volume of user-generated textual data is increasing every day. Therefore, with the rapid increase of online reviews, it is impossible for people to review all comments related to a product or service in a limited time [3]–[5]. To alleviate this problem, Amazon and some other online retailers allow consumers to evaluate opinions by implementing useful voting systems. In these systems, beside each comment, there is a question such as: ''Was this comment helpful to you?''. Also, usefulness information is usually reported in the form of a usefulness score to help the consumer evaluate the review [3]. This usefulness score is equal to the ratio of the number of useful votes to the total number of votes for a given opinion and is written as ''n out of t people found this helpful'' [6].

The associate editor coordinating the review of this manuscript and approving it for publication was Shiqing Zhang.

Although the usefulness score calculated on the basis of user votes may be informative, it suffers from the ''cold-start'' problem. In other words, it cannot be used properly for recent (newly written) reviews, having not yet been voted on. Hence, due to a lack of information, computing the usefulness score of such newly written reviews is more complicated [7], [8]. Another problem is that the comments that are posted earlier attract the most readers' attention, and are consequently placed at the top of the list, while comments that are posted later are listed at the bottom of the list and ignored [4], [5]. This phenomenon is known as the Matthew effect [9]. Another problem with usefulness score is that as reviews are quickly posted, some useful reviews are likely to be covered by useless reviews before being considered [8]. Therefore, it is important to use content-based methods for automatically analyzing reviews that can accurately predict the usefulness score of reviews as soon as they are posted on a website.

The importance of automatically recognizing useful comments has been examined in previous studies [1], [4]–[6], [10], [11]. However, most existing studies have some limitations. For example, few studies have examined the effect of various emotions such as sadness, happiness, trust, expectation, etc. on the usefulness of reviews though, many researchers have acknowledged the role of emotion in the online environment and emphasized the importance of examining the role of emotion in interpreting online reviews [6], [10]. Moreover, it has been shown that emotions from the same polarity (e.g., anger and fear) may have different effects on consumers' activities including their perceived usefulness of a review [10]. Consequently, it is necessary to consider the effect of different emotions about the usefulness of reviews. Another problem of the existing methods is that they do not consider the intensity of the words in each emotion category. This is important because it is possible for two different words to represent the same emotions with different intensities. For example, the two words ''good'' and ''great'' both convey a sense of pleasure, but it is clear that each convey a different intensity from that emotion. To our knowledge, previous studies on review usefulness prediction did not addressed this problem. In the current study, this issue is considered when extracting an emotion from a review. Also taking into account the three semantic dimensions of valence, arousal, and dominance (VAD) along with the dimensions of emotion for the context of each review can lead to a more detailed analysis because, according to [12], analyzing these three dimensions is very effective in understanding the meaning of the text. It is also expected that the more accurate the meaning of the review is, the more useful it will be.

In addition to the abovementioned problems, some previous studies' results need to be investigated more carefully. For example, the results of previous research indicate that surprise and expectation do not play a critical role in the classification of usefulness of reviews [10]. And the two emotions of surprise and anticipation fall into the category of positive emotions. However, examining the vocabulary of these emotions shows that they are two-sided (i.e., they may be positive or negative). In fact, positive surprise and expectation can produce positive opinion while negative surprise and expectation can lead to negative opinion. These have not been considered in previous studies on review helpfulness prediction. Therefore, in this study, each of these two emotions is divided into two groups of positive and negative. This, increases the diversity of existing emotion categories in the NRC lexicon [13], [14] (i.e., sadness, pleasure, fear, anger, hatred, trust, surprise and expectation) to 12 by adding four ''positive surprises'', ''negative surprises'', ''positive anticipation'', and ''negative anticipation'' emotions.

Having extracted the above features, to help avoid biasing the review helpfulness classifier, we train seven classifiers and selected top three ones to combine their results. Specially, each of three classifiers assign a probability to each usefulness category. These probabilities may be interpreted as the confidence level of classifiers in assigning categories to reviews. To obtain the final helpfulness category of reviews we proposed an improved evidential fusion method. This method is based on the Dempster-Shafer (D-S) theory which has been studied extensively in decision-making and sentiment analysis in recent years [15], [16]. This algorithm is used to fuse information obtained from different sources, especially when there is uncertainty in the sources [16]–[18].

Considering the mentioned aspects of the current study, in the proposed model, we seek to answer the following questions to determine the usefulness of reviews:

1) Does considering three dimensions of valence, arousal, and dominance (VAD) along with contextual and emotion-related features affect the usefulness of the comments?
2) Does the improvement of the D-S algorithm increase the accuracy of the review usefulness prediction system?

The main contributions of this study are as follows:

1) Given that both surprise and expectation in the NRC emotion lexicon can have either negative or positive polarity, the NRC lexicon is enhanced by adding four emotions namely, positive surprise, negative surprise, positive expectation, and negative expectation.
2) A new algorithm for extracting 12 separate emotion vectors from the reviews is presented that considers the following aspects:
    a. To increase the accuracy of the extracted emotion, the intensity of the emotion of words in separate emotional groups is also considered.
    b. To extract the emotion vector correctly, the effect of negative structures in the sentence has been investigated.
    c. To extract the emotion vector more accurately, the effect of intensifying and decreasing emotion structures on words has also been investigated.
3) Three features of valence, arousal, and dominance (VAD) for each text for detecting the helpfulness score of reviews is considered.

4) To determine the usefulness of the reviews, the triplet structure is used to improve the D-S algorithm.

The findings of this study can be used to improve e-shops, opinion mining, review summarization, and recommender systems. Also, the emotion and sentiment extraction parts of the proposed system may be used separately in sentiment and emotion analysis.

The remainder of the paper is organized as follows. In Section II, related work is presented, Section III presents the proposed method in more details. Section IV analyzes the results of the implementation, then discusses the findings. Section V concludes the paper and presents some directions for future work.

II. RELATED WORK

A. EXISTING STUDIES ON THE USEFULNESS OF REVIEWS

The usefulness of a review means the objective evaluation of the quality of the review by others [10]. Consumers can hardly manually identify useful reviews among the high volume of product reviews on websites, so finding useful reviews automatically and understanding of the factors influencing the usefulness of reviews are important [6]. The usefulness of online reviews is a multi-dimensional concept that can be controlled by a variety of factors [10]. Many researchers have suggested various factors that may influence the usefulness of reviews. For example, review length that is the number of words that constitute the review may be an influential factor [1], [19]–[23]. Several studies have shown the positive effects of review length [1], [21], [22], [24], [24]–[29]. This effect varies with regard to product type; review length has a greater effect on the usefulness of search product than experimental goods [1], [27].

In 2016, Qazi *et al.* [30] presented a conceptual model for predicting the usefulness of reviews. This study not only examines quantitative factors (such as the number of concepts), but also focuses on the qualitative aspects of reviews including types of review (i.e., regular, comparative, and suggestive reviews). A comparative review expresses a relationship of similarity or difference between two or more entities, or describes an individual's preference for common features of entities. Normal opinions express a general view and suggestive reviews give advice on whether or not to buy a product [30]. The researchers found that all three types of reviews had significant effects on the purchase decision [4]. There are different opinions about the impact of subjectivity/objectivity on the usefulness of reviews. Facts contain objective statements about their entities, events, and attributes while subjective expressions describe individuals' emotions and evaluations of their entities, events, and attributes [31]. The effect of subjectivity on the usefulness of opinion has been demonstrated in [8], [32]–[35].

Researchers also examined several aspects of review text such as various readability measures, spelling errors, subjectivity levels, and average usefulness of reviews [33]. It has been shown that opinions with a combination of objective and subjective sentences have a negative relationship with product sales compared to only subjective or only objective ones [33]. The researchers also found that mid-length reviews with a few misspellings were more effective from customers' view than very long or very short reviews with more misspellings [4], [33]. However, in [20] the authors concluded that legibility is more effective than review length [4]. In [8], the characteristics of noun, verb, and adjective have been used as effective predictors and in [6], the adverb feature is also defined alongside the noun, verb, and adjective. Linguistic features are also used to predict usefulness [35].

Several studies have shown that the readability of the review text is an effective factor in predicting the usefulness of the review [6], [8], [35], [36]. Reviews with high readability are likely to be read and receive more votes from users [6]. In addition, previous studies have shown that total votes and total number of reviewers are also effective features of predicting the usefulness of online reviews [37]. A new reviewer feature, namely reviewer's activity length along with total votes and total number of reviews are also proposed [6]. Also disclosed personal information such as real name, location, nickname is shown to be effective features to predict the usefulness of online reviews [38].

In [8], a few models were designed to predict the usefulness of consumer reviews using several contextual features such as polarity, subjectivity, entropy, and ease of reading. These machine learning models automatically determine the usefulness of each initial review as soon as it is posted on the website so that they have an equal chance of being viewed by others. Authors in [39] examine how online consumers' reviews interact with one another and how consumers' beliefs evolve over time. The researchers proposed a dynamic model of opinion evolution that is applicable to the opinions of online consumers in the e-commerce environment and influencing factors such as visitor readability, sorting and dissemination strategies, convergence parameters, feedback, and thresholds of trust. In [40], a review usefulness prediction framework is proposed to use multilingual reviews to generate relevant business insights and predicts the usefulness of reviews with the help of non-English comments.

Previous studies show that product features also play an important role in predicting the usefulness of online reviews [1], [6], [37]. Types of products (i.e., experimental and search), can play an important role in the usefulness of reviews [6], [41]. Determining the quality of empirical products before use is not easy, so consumers looking for

empirical products use others' experiences, but search products can be judged on the basis of product specification before purchasing [6], [27], [42]. The researchers found that positive and negative emotions in search products were more effective than empirical ones. In addition, for search products, the combination of features with positive and negative emotions performs better than experimental goods [6].

In [6], four product features namely, product emergency index, Amazon sales rank, Amazon product price list, and time spent from product release date were proposed. In [5], product description and question-answer features along with contextual features are used. The results show that using product description features and customer question-answer data improves the accuracy of predicting usefulness scores. The characteristics of product reviews among five different products were investigated in [43] and their effects on the usefulness of the review was identified. Four data mining methods have also been explored to determine the best way to predict the usefulness of comments for each product using five Amazon datasets. The results show that opinions for different types of product have different linguistic and psychological characteristics and their influencing factors are different.

The model presented in [44] assumes that product feedback sends signals to buyers. Using the Amazon product reviews, the researchers tested their model and found that signals related to comment content (e.g., specific comment content and writing style) and comment-related signals (such as reviewer experience and his/her popularity) are both influencing review usefulness. In addition, they showed that the signaling environment was influenced by the signal and that the motivations given to the reviewers influenced the signals sent.

B. EMOTIONAL FEATURES AND REVIEW USEFULNESS

Several psychological theories have suggested basic human emotions of various dimensions [45]–[47]. In [45], six emotional dimensions are suggested: pleasure, sadness, anger, fear, disgust, and surprise. Plutchik showed eight views of emotion, sadness, pleasure, fear, anger, disgust, trust, surprise, and anticipation using a wheel [48]. The Plutchik framework has strong foundations in psychological studies [10] and unlike some other models [45] in which negative emotions are predominant, in this framework, the balance between negative and positive emotional perspectives is established.

Several studies investigated the effect of different emotions on the usefulness of reviews. For example, researchers in [1] examined the relationship between few negative emotions (anger, fear, sadness) and the usefulness of online reviews. The results of this study showed that fear in a review has a positive effect on its usefulness and anger has a stronger negative effect on the perceived usefulness of the review for experimental goods than search goods. As the level of sadness in reviews increases, the perceptual usefulness of the review decreases.

It has been shown that the effect of negative consumer reviews on film choice was more effective than positive ones [49]. In contrast, positive reviews have a greater impact on film evaluation than negative ones. Also, consumer expectations have been a moderating effect of the capacity of consumer opinion on film selection and subsequent evaluations. Similarly, in [50], a statistical approach was proposed to establish the relationship between review quality and emotions with review usefulness. The results showed that high quality positive reviews increase product sales compared to low quality reviews [4]. In [6], the effect of four positive and four negative emotions on perceptual usefulness is investigated and a binary classification model is developed to predict usefulness based on deep neural network. In addition, product type features, visibility, readability, linguistic and sentiment characteristics of reviews are also used to compare and predict the usefulness. Experimental results showed that positive emotional traits perform better. However, negative emotions and visibility are also affected. Also, a combination of features with positive emotional traits provide the best performance for online feedback. Also, trust, pleasure, anticipation (positive emotions), anxiety, and sadness (negative emotions) are the most effective emotional dimensions and have a greater effect on perceptual usefulness.

Emotional content was assessed with eight emotional dimensions (pleasure, sadness, anger, fear, trust, hatred, expectation, and surprise) in the Plutchik emotional wheel [10]. The results of this study were analyzed using a negative binomial model which showed that anger, hate, and fear in reviews had a positive effect on the usefulness of the reviews, and pleasure, trust, and sadness had negative effects on the usefulness of the reviews. The anticipation and surprise in the reviews had no effect on the usefulness. In [11], the influence of emotions in the helpfulness of hotel-related online reviews is examined. The results of this study showed that negative online reviews are more useful than positive ones. It has been also shown that quantitative and capacity-based approaches are not sufficient to identify and evaluate the quality and performance of hotel services in information seeking and decision-making processes for consumers.

Table 1 depicts the summary of literature and major determinants of perceived review usefulness in studies.

C. SCORE FUSION METHOD USING AN EVIDENTIAL APPROACH

One way of improving the efficiency and accuracy of machine learning systems is to fuse the results of various classifiers [51]. This can be achieved through the use of score fusion algorithms [51]. In the existing review usefulness systems, most employed fusion methods have no theoretical basis and not designed for these systems. To address this problem, we exploit an evidential fusion method based on the Dempster-Shafer (D-S) theory. D-S method is one of the most prominent score fusion methods which has been exploited in recent years for polarity detection [52], rating prediction [15], multimodal emotion recognition [53], and project risk assessment [54]. In terms of uncertainty in the validity of the hypotheses, Dempster and Shafer presented a general form of Bayesian theory in which multiple probabilities (e.g., derived from multiple classifiers' outputs) were used to determine the final output on the basis of evidence from uncertain outputs [51]. Using the D-S theory, evidences are first extracted from the classifiers' outputs. These evidences are then used as basic knowledge in finding the degree of membership of the input to each class. Based on this evidence, the probability

TABLE 1. Summary of literature from 2008 to 2020 on review usefulness.

Category	Feature name	Definition	Author
Content	Length of the review	The number of words in the review.	Ren and Hong [1], Gao et al [19] , Korfiatis et al [20], Kuan et al [21], Siering and Muntermann [22], Weathers et al [23], Yin et al [24], Baek et al [25], Peng et al [26], Mudambi and Schuff [27], Salehan and Kim [28], Kuan et al [21], Schindler and Bickart [29], Qazi et al [30]
Review type	Comparative, Regular and suggestive	A comparative review expresses a relationship of similarity or difference between two or more entities, or describes an individual's preference for common features of entities. Normal opinions express a general view and suggestive reviews give advice on whether or not to buy a product.	Qazi et al [30]
	Objective/Subjective	the review contains the views and evaluations of its writer or just presents some facts.	Singh et al [8], Indurkhya and Damerau [31], Ghose et al [32], Ghose and Ipeirotis [33], Liu et al [34], Krishnamoorthy [35]
Linguistic	Noun	Percent of nouns in the review.	Saumya et al [5],Malik and Hussain [6], Singh et al [8], Krishnamoorthy [35]
	Verb	Percent of verbs in the review.	Saumya et al [5], Malik and Hussain [6], Singh et al [8], Krishnamoorthy [35]
	Adverb	Percent of adverbs in the review .	Malik and Hussain [6], Krishnamoorthy [35]
	Adjective	Percent of adjectives in the review.	Saumya et al [5], Malik and Hussain [6], Singh et al [8], Krishnamoorthy [35].
	Pronoun	percentage of words in the review text that are pronouns	Malik and Hussain [6]
	Article words	percentage of words in the review text that are article words	Malik and Hussain [6]
	Prepositions	percentage of words in the review text that are preposition	Malik and Hussain [6]
Textual	Readability	simplicity of understanding the text by the user	Saumya et al [5], Malik and Hussain [6], Singh et al [8], Krishnamoorthy [35], Korfiatis et al [20], Hu and Chen [36].
Context	Rating	Review Rating (from 1 to 5 stars)	Ren and Hong [1], Malik and Hussain [6], Lee and Choeh [37], Forman et al [38]
	Reviewer disclosure	disclosed personal information by reviewers such as: real name, nickname and location	Forman et al [38]
Reviewer	Activity Len	Time between first review and last review	Malik and Hussain [6]
	Reviewer votes	Reviewer total votes	Malik and Hussain [6]
	Reviewer reviews	Reviewer total reviews	Malik and Hussain [6]
Product	experimental and search	The recognition of the quality of experimental products is not easy before use, so consumers who use experimental products are using others' experiences, but search products can be judged on the basis of product specification before purchasing	Ren and Hong [1], Malik and Hussain [6], Lee and Choeh [37], Mudambi and Schuff [27], Pan and Zhang [41],Willemsen et al [42], Park [43]
	product description	Similarity between product description and review text	Saumya et al [5]
	Price	Product price (Low vs. High)	Malik and Hussain [6], Baek et al [25]
Psychological	Drives words ,Space	percentage of words in the review that	Malik and Hussain [6]

TABLE 1. ***(Continued.)*** **Summary of literature from 2008 to 2020 on review usefulness.**

		are drive words. percentage of words in the review that are Space words.	
Emotions	Fear Sadness Disgust Anger	Negative Emotions	Ren and Hong [1], Malik and Hussain [6], Wang et al [10], Lee et al [11]
	Joy Trust Surprise Anticipation	Positive Emotions	Malik and Hussain [6], Wang et al [10], Lee et al [11],
Sentimental	Sentiment	Sentiment of review text	Malik and Hussain [6]
	Polarity	Polarity of reviews (positive, neutral, negative)	Malik and Hussain [6], Baek et al [25], Pan and Zhang [41],Tsao [49], Lee and Shin [50]
	intensity	emotional intensity	Peng et al [26]

masses of each class are determined. Finally, the masses are fused to determine the final class [51].

In [51], a D-S theory-based approach for combining multiple classifiers was designed and the class outputs are modeled by a 2-point Triplet structure. The results showed that the first and second best classifier performs better than the separate classifiers and hybrid classifiers. The triplet structure also performs better than the simplet and quartet structures. In [55], a hybrid method for text classification was designed based on the D-S theory that used combination of best outputs. This method also used the outputs of the classifiers using the quartet or 3-point structure. They compared the performance of separate classifiers with the hybrid SM (the combination of SVM and KNNM) classifier. The researchers found that the best hybrid classifiers had higher performance than the best independent ones.

The original D-S theory is used in [18] to design a method to detect sentiment in online reviews. The researchers showed that the D-S theory-based fusion method performs better than simple fusion methods such as weighted average and sum using both lexicon-based and machine learning-based methods. They improved their method for detecting document-level review polarity in [17]. In this study, the TripAdvisor and CitySearch datasets was used to evaluate the performance of the improved D-S-based fusion method. The researchers found that their proposed method was more accurate than the original D-S fusion method.

In [56], a D-S theory-based approach was designed to combine conflicting evidence with different weighting coefficients and provide a high-performance decision support system that can effectively solve the collision problem. In [57], a method based on the D-S theory was designed that used absolute and relative difference factors for two pieces of evidence. The resources were divided into two categories of collision and non-collision, and the cumulative probabilities in the collision category was combined with those of the non-collision group. The advantages of this method are better management of evidences and improving reliability.

In [58], a new method is proposed to incorporate social media comments with audio-visual contents in video. For the fusion stage, the decision-level fusion method was used based on the D-S theory of evidence. The results showed that the D-S-based fusion system performs much better than the baseline method, which uses only audio-visual content for emotional video retrieval. In a similar study [59], a new lexicon-based method for information fusion based on D-S theory was proposed. This method does not require a human-coded corpus for training and operates much faster than the supervised method. The results showed that inclusion of song lyrics with audio-visual content had no positive effect on the retrieval performance, but utilizing users' comments had a significant improvement for the emotional retrieval system. To address the main drawback of combining multimodal information in emotional video retrieval systems which is assigning equally weights to modalities, a new D-S method was proposed in [60]. This method gives different weights based on the correlation and the level of confidence. This method has been recently improved for the same task using a hybrid architecture consisting of latent information obtained through canonical correlation analysis (CCA) [61] and Marginal Fisher Analysis (MFA) [53]. As stated in some of previous studies, the original D-S theory has some limitations [51]; One of the most influential limitations is the production of contradictory results. To solve this problem, the triplet structure is used to improve the D-S algorithm [51].

III. THE PROPOSED METHOD

The overall view of the proposed model is shown in Figure 1. In this system, reviews are first preprocessed. Then, using the proposed emotion extraction algorithm, the emotion-related features are extracted by considering the emotion intensity of

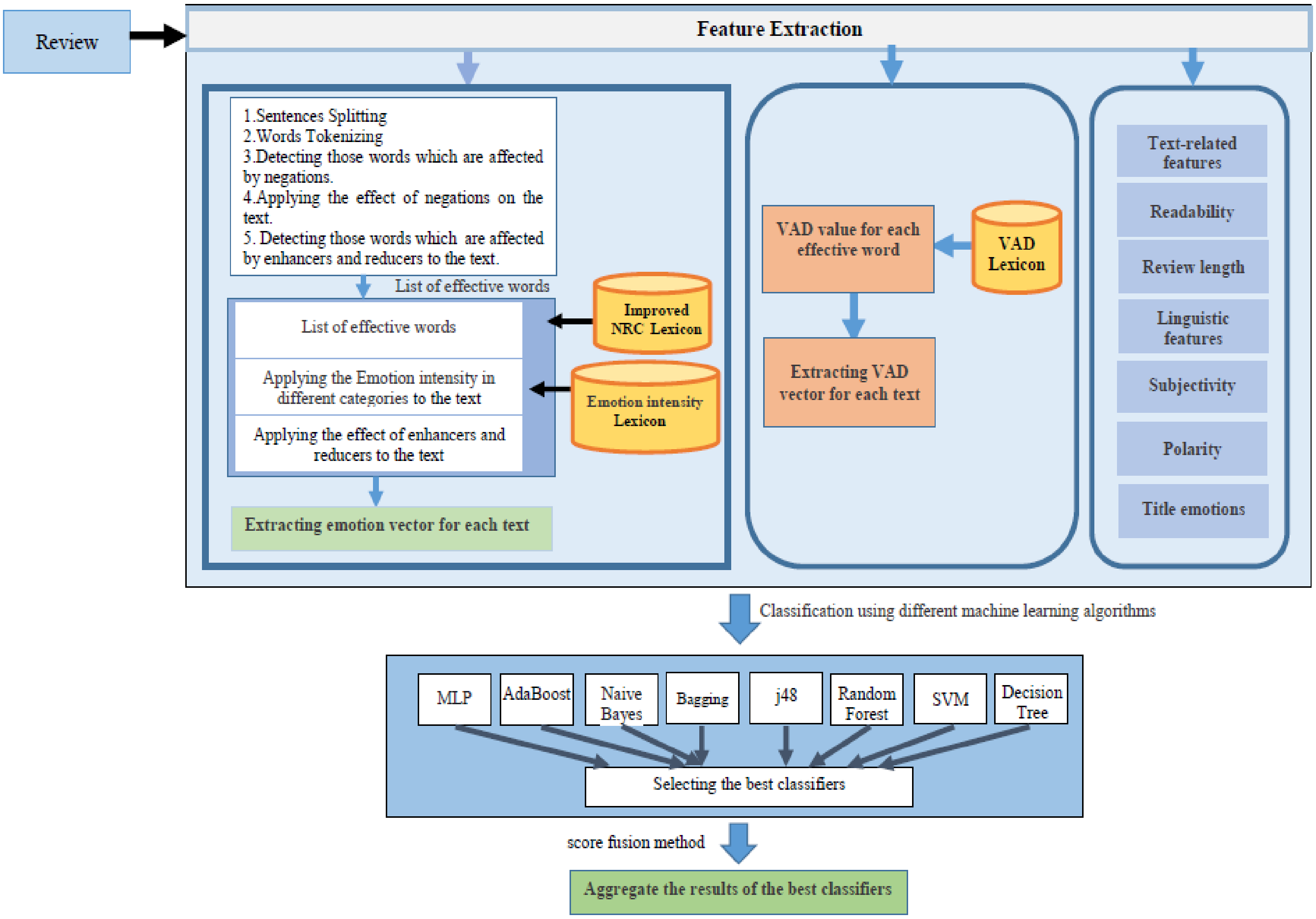
Review
Feature Extraction
1.Sentences Splitting
2.Words Tokenizing
3.Detecting those words which are affected by negations.
4.Applying the effect of negations on the text.
5. Detecting those words which are affected by enhancers and reducers to the text.
List of effective words
List of effective words
Applying the Emotion intensity in different categories to the text
Applying the effect of enhancers and reducers to the text
Improved NRC Lexicon
Emotion intensity Lexicon
Extracting emotion vector for each text
VAD value for each effective word
VAD Lexicon
Extracting VAD vector for each text
Text-related features
Readability
Review length
Linguistic features
Subjectivity
Polarity
Title emotions
Classification using different machine learning algorithms
MLP
AdaBoost
Naive Bayes
Bagging
j48
Random Forest
SVM
Decision Tree
Selecting the best classifiers
score fusion method
Aggregate the results of the best classifiers

FIGURE 1. The overall view of the proposed model.

the words in different emotional categories and the affective shifters. Additionally, other features such as linguistic features, readability, valence, arousal, dominance values, review length, and polarity are also extracted for each review. In the next step, using different machine learning algorithms, different models are developed to predict the usefulness of reviews and the best three models are selected based on the model evaluation criteria. Finally, to improve the learning accuracy, the results of the best classifiers are fused using the D-S fusion algorithm. The proposed method is explained in more details in the following sections.

A. PROBLEM FORMULATION

Assume R is the initial unprocessed dataset of size n×3 where n is the number of reviews in R and the columns contain review text, title, and helpfulness score of the review. The goal is to predict the helpfulness score $s \in \{0, 1\}$ of a given test review r using the proposed classification method trained on R. To this aim, the problem is first reduced to extracting k representative features from the first and second columns of R then, to use the proposed method to classify reviews. The feature matrix F of size $n \times k$ is constructed by extracting review text and title, different features such as emotion vectors, linguistic features, and other derivable features from these two.

B. FEATURES

The final features extracted from the initial dataset are presented in Table 2. In the following section, each of these features is described in detail.

C. EMOTION VECTOR EXTRACTION FROM EACH REVIEW

The NRC emotion lexicon [13], [14] contains 14,183 words that provide a distinct emotion for each of the 8 dimensions. In this study, four positive emotions of positive surprise, negative surprise, positive expectation, and negative expectation were added to the NRC lexicon. If a word had a surprise emotion, it would have been considered a positive surprise if it had positive polarity, otherwise it would have been considered as negative surprise. In case the surprised word had no polarity, these two new feelings would have been assigned zero. Also, if the word did not have a surprise emotion, the two new emotions will be assigned zero. This will be performed similarly for the expectation emotion. Ultimately, an upgraded dictionary contains 12 distinct emotions for each word. Also, since many of the words in the NRC lexicon have

TABLE 2. Features used in this study to train machine learning algorithms for predicting product review helpfulness.

Feature name	Feature category	Description
Emotion vector	Emotion-related	A vector containing of 12 elements for 12 distinct emotions, each representing the degree of emotion involved.
Title's emotion		The emotion extracted from the comment title text.
Polarity of the review		Polarity of the review (positive, negative)
OneLetterWords	Text-related features	The number of one-letter words in the review [8].
TwoLetterWords		The number of two-letter words in the review [8].
LongerLetterWords		The number of more than two-letter words in the review [8].
VAD		Vector consisting of 3 elements corresponding to V, A, and D values.
Readability		The degree of simplicity of understanding the text by the user. The textstat library in Python was used to extract this feature in range [0-100] as follows[8]: *90-100 :Very Easy *80-89 :Easy *70-79 :Fairly Easy *60-69 :Standard *50-59 :Fairly Difficult *30-49 :Difficult *0-29 :Very Confusing
Subjectivity		Whether the review reflects the opinions and evaluations of the user. Using the TextBolb library, the subjectivity of the sentence is calculated from zero to one, and the average subjectivity of these sentences is calculated as the amount of subjectivity of the text
Review length		Number of words forming the review.
Name		Percent of the names in the review.
Verb		Percent of the verbs in the review.
Adverb		Percent of the adverbs in the review.
Punctuations		Percent of the punctuations in the review.

zero values for all of their separate emotions and hence have no effect on the final feature vector, these words have also been removed from the improved dictionary. This changed the size of the improved NRC lexicon to 6469 distinct words. The details of the improved lexicon are listed in Table 3.

To improve the accuracy of computing the perception of review texts, the NRC Affect Intensity Lexicon was also used in this study [62]. Specifically, for each review, a 12-elements vector is extracted each of which is equivalent to the amount of emotion in a separate emotional dimension. This vector can be expressed as $EV = $ (*Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, Trust, Positive-Surprise, Negative-Surprise, Positive-Anticipation, Negative-Anticipation*). To calculate this vector for each review, after preprocessing using Algorithm 1, affective words are extracted from the text. Then, using Algorithm 2, the emotion

TABLE 3. Specifications of the improved NRC lexicon.

Emotion	Word count
Anger	1247
Anticipation	839
positive Anticipation	449
negative Anticipation	143
Disgust	1058
Fear	1476
Joy	689
Sadness	1191
Trust	1231
Surprise	534
positive Surprise	198
negative Surprise	196
Total	9251

vector is extracted for this set of words which is equivalent to the whole emotion vector of the text being processed.

To correctly extract the sentiment of the text, it is necessary to identify the emotionally effective words as accurately as possible. In this study, the following preprocesses were made on the text of the reviews.

1) THE EFFECT OF NEGATION

In any linguistic structure, negation words such as "not" may be used to reverse the emotion of words with a positive affective sense [63]. In this study, Python Spacy library was used to identify the words that were negatively affected. Then, each word was replaced with its best antagonist using WordNet, finally, the negation word was removed.

2) INFLUENCE OF QUALIFIERS ON TEXT

Qualifiers in English reduce the impact of the word they are directly associated with. For example, in the term "hardly crashes", the word "hardly" reduces the semantic impact of the verb "crashes", and if the verb has a negative affect, it is more appropriate to reduce the negative effect by considering the effect of the word "hardly". The list of words of this category considered in the current study is as follows:

Qualifiers = ['hardly', 'rarely', 'infrequently', 'seldom', 'sporadically', 'scarcely']

To identify words that are affected by qualifiers in the review text, when examining the negative relation for each word, if the word has an adverb modal relation with another word and it (i.e., the original word) has a POS tag of "pronoun", it is checked whether it is in the list of qualifiers or not; if any, a word that is directly affected is identified and added to the list. The relationships and POS tags are also extracted using the Python Spacy library. Finally, to reduce the effect of emotion in practice, when calculating the emotion, a constant value of -0.2 is added to the emotional value of each affected word.

3) THE EFFECT OF POSITIVE INTENSIFIER IN THE TEXT

In English there are words that if they come before another word, increase the semantic effect of the word affected. for example, in the phrase "very good", the word "very" intensifies the semantic effect of the word "good". In this study, the following words are considered in the intensifier list:

Positive intensifiers = ['very', 'extremely', 'absolutely', 'completely', 'greatly', 'too', 'so', 'totally', 'utterly', 'highly', 'rather', 'really', 'exceptionally', 'particularly', 'seriously]

To exaggerate the emotion value of the words affected by the intensifiers in the review text, a similar approach described for qualifiers is done except that the words affected by this list are in a separate list and when the emotion value is calculated, a constant value of 0.2 is added to the emotional value of each affected word [63].

4) THE EFFECT OF BOTH POSITIVE INTENSIFIERS AND NEGATIONS

If a word is simultaneously affected by both negative and positive influencers, instead of increasing the emotion value, it should be reduced. For example, to extract the correct emotion value of the word "not very good", the word "good" as mentioned before is replaced by its opposite word (i.e., "bad"). Then, the intensifier word "very" is considered as qualifier and reduce the emotion value by -0.2.

According to Algorithm 1, first all possible relationships between the two words in the sentence are extracted by the Spacy library in Python (line 7). Then, in lines 8-11, if there is a negative relationship between the two words w_i and w_j (i.e., the word w_j is affected by the negative word w_i), it is checked whether the word w_j is also affected by the word intensifier w_z and if it is affected, it is added to the list of words affected by the diminished words. Then, according to lines 13-14, the word w_j is replaced by the WordNet opposite of w_j and the word w_i is deleted.

In lines 16-19, if the relationship between the two words is an adverb modal relation and the role of the word w_i is intensifier, it is first examined whether the word w_i is included in the list of intensifier words. If w_i exist, wj is added to the list of words affected by the intensifier. If the word w_i is included in the list of reducer words, the word affected by w_j is added to the list of words affected by the diminished words (lines 20-21). Finally, lines 25-29 are added to the final word list by examining each word in the sentence s_i if these words are other than stop words and presented as the output of the algorithm.

The objective of Algorithm 2 is to extract the emotion vector (EV) for each review u_i. This vector has 12 elements,

Algorithm 1 Pre-Processing to Extract Effective Emotional Words

```
1:  Input: one sentence from review $u_i$.
2:  Intensifiers list = [very, too, ...]
3:  Qualifiers list = [hardly, rarely, ...]
4:  Intensified words []
5:  Diminished words []
6:  finalWords []
7:  Relations = extract relations between $s_i$ words ($w_i$,$w_j$)
8:  For each relation in relations do
9:    If relation is negation relation then
10:       If $w_j$ have adverb modal relation with $w_z$
        And $w_z$ is in Intensifiers list then
11:         Add $w_j$ to Diminished words
12:       End if
13:       Opposite affected word $w_j$ using wordnet in $s_i$
14:       Delete negated word $w_i$
15:     End if
16:     If relation is adverb modal and $w_i$ pos is
        ADV then
17:     If $w_i$ is in Intensifiers list then
18:         Add $w_j$ to Intensified words
19:     End if
20:     If $w_i$ is in Qualifiers list then
21:         Add $w_j$ to Diminished words
22:     End if
23:     End if
24:   End for
25:   For each $w_i$ in si do
26:     If $w_i$ is not stopWord then
27:       Add $w_i$ to finalWords
28:     End if
29:   End for
```

each corresponding to a separate emotion. Therefore, each element of this vector indicates how much emotion of the corresponding dimension is in the review. To extract the EV_{ui}, first break the sentence and then execute Algorithm 1 for each sentence. This provides a list of the effective words of the sentence, as well as the words that are affected by the intensifier and reducers (Lines 1-15). Then, for each word in the list, the EVw_i vector, which represents the values of the individual emotions for the word, should be extracted. This will be done using the values available for the word or, if such values do not exist, the values for the word lemma in the NRC emotion lexicon, as well as the values of the intensity available for the word, or if such values do not exist, the values available for the word lemma in the NRC affect lexicon (lines 16-21).

calculatedEmotionEffectForWord function takes each word and its emotion vector and adds corresponding values of emotion intensity in different groups to emotion vector of the word. If the word or its lemma is not present in NRC lexicons, its existence in each of the affect dictionary groups is checked and its values are added when values are found (lines 22-30).

Algorithm 2 The Proposed Emotion Extraction Algorithm

```
1:  Input: Collection of all reviews ($U$)
2:  For each review $u_i$ in $U$ do
3:    $EVu_i$ = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]; //
      $EV_i$ = emotion vector
4:  For each sentence $s_i$ in $u_i$ do
5:    $s_i$ final words = preprocess($s_i$);
6:    $EVs_i$ = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0];
7:    For word $w_i$ in $s_i$ final words do
8:    Intensify value = 0;
9:    Diminishing value = 0;
10:     $W_i$_lemma= lemma of $w_i$;
11:     If $w_i$ in Diminished words then
12:     Diminishing value = -0.2;
13:     Else if $w_i$ in Intensified words then
14:       Intensify value =0.2;
15:     End if
16:     If$w_i$ in NRC_Emotion_Lex then
17:       $EVw_i$ = read row $w_i$ from NRC_Emotion_Lex;
18:       calculateEmotionEffectsForWord($w_i$,
        $EVw_i$);
19:     Else if $w_i$_lemma in NRC_Emotion_Lex then
20:       $EVw_i$ = read row $w_i$_lemma
        from NRC_Emotion_Lex
21:       calculateEmotionEffectsForWord
        ($w_i$_lemma, $EVw_i$);
22:     Else
23:       For each emotion e in NRC_Affect_Lex do
24:         If$w_i$ in NRC_Affect_Lex and group e then
25:           - $EVw_i$[e] + = value of
          NRC_Affect_Lex for ($w_i$,e)+1;
26:         Else if $w_i$_lemma
          in NRC_Affect_Lex and group e then
27:           $EVw_i$[e] + = value of NRC_Affect_Lex
            for ($w_i$_lemma,e)+1;
28:           End if
29:         End for
30:       End if
31:       For each non zero element x in $EVw_i$ do
32:         $EVw_i$[x]+ = Intensify value + Diminishing
          value;
33:       End for
34:       $EVs_i$+ =$EVw_i$;
35:     End for
36:     $EVu_i$+ =$EVs_i$;
37:   End for
38: End for
```

Also, if the word is affected by any intensifier or reducer, then the values of non-zero EVwi elments decrease or increase steadily (lines 31-33). These values are then summed for all sentence words to make the EVsi, which is the emotion vector for the sentence. Finally, all EVsi vectors are summed to form EVui vector (lines 34-36).

calculateEmotionEffectsForWord

```
calculateEmotionEffectsForWord
  (word w,EmotionVector EVw[]){
 **For** each emotion e in NRC_Affect_Lex **do**
**If** w in NRC_Affect_Lex and group e **then**
EVw [e]+ = value of NRC_Affect_Lex for (w,e);
**Else if** w_lemma in NRC_Affect_Lex and group e
  **then**
EVw [e]+ = value of NRC_Affect_Lex for
  (w_lemma,e);
 **End if**
**End for**
  }
```

Algorithm 3 Extracting VAD Vector for Each View

```
1:  For each review u_i in U do
2:      VAD _u_i = [0, 0, 0]; // VAD_u_i = VAD vector
3:      sentCount = 0;
4:      For each sentence s_i in u_i do
5:          VAD _s_i = [0, 0, 0];
6:          sentCount ++;
7:          wordCount =0;
8:          For each word in s_i word_list do
9:              If word is in VAD _Lex then
10:                 wordCount ++;
11:                 VAD _ s_i+ = value of [V,A,D] for word;
12:             End if
13:             End for
14:         VAD _s_i = VAD_s_i / wordCount;
15:         End for
16:         VAD _Vu_i+ = VAD _s_i;
17:     End for
18: VAD _Vu_i = VAD _Vu_i/ sentCount;
```

D. EXTRACTING THE VAD VECTOR OF EACH VIEW

According to [12], in addition to the emotional dimensions that are transmitted through words, the three semantic dimensions of valence, arousal, and dominance (VAD) are also transmitted through the words of a text. The *v* (positive or negative/pleasant or unpleasant) dimension is a measure of whether or not a word is favorable. For example, the word "party" indicates a higher level of positive than "funeral". The *A* (irritability/not irritability) dimension measures how energetic or crooked it is felt and this is not a measure of emotion intensity. Sadness and depression can be low irritations and severe emotions. While anger and wrath are unpleasant emotions, they have higher irritability than laziness. The dimension *d* (dominant-submissive and subordinate) represents the sense of being obedient and dominant. For example, the "battle" is more dominant than "delicate".

Considering these three semantic dimensions along with the dimensions of emotion for the context of each review can lead to a more detailed analysis. The analysis of these three dimensions is very effective in understanding the meaning of the text [12]. As to the usefulness of comments, it is also expected that the more accurate the meaning of the review is, the more useful it will be. Algorithm 3 extracts the VAD value for each text.

To extract the VAD vector of each text after extracting the effective word list, the existence of each of these words in the VAD dictionary is checked and, if any, the VAD vector is extracted for each word (Lines 1-10). Finally, by averaging over these vectors, the VAD vector of each sentence is obtained in the text (Line 11-15). Finally, to extract the VAD vector of the text, it is sufficient to carry out the averaging text sentences on the VAD vectors (Lines16-18).

The process of pre-processing and extracting the emotion vector for the text in Example 1 is as follows:

Example 1: "it is not good, beautiful and very cold but very delicious and rarely uses".

1) PRE-PROCESSING

1. Determine qualifier affected words: First, it is determined that the word "uses" in the sentence is affected by a reducer.

Qualifier affected: ['uses']

2. Determine positive intensifier affected words the words that were affected by the intensifiers were included in the positive intensifier affected list:

Positive intensifier affected [cold, delicious]

3. Determine negative affected words all the words that have been negatively impacted are listed in the negative affected words list:

Negative affected words [good, beautiful, cold].

As can be seen, the words that are simultaneously negatively affected and intensified can be obtained by intersecting the two lists of positive intensifier-affected and negative-affected words.

Finally, after replacing the negatively affected words with their antagonisms and extracting the correct effective words, the following is a final list of preprocessing algorithm outputs.

Word list: ['good', 'beautiful', 'hot', 'cold', 'uses']
Final list: ['bad', 'ugly', 'hot', 'delicious', 'uses']

2) EXTRACTING EMOTION VECTOR

Emotion extraction algorithm applied on existing words in final list to extract emotion vector in example 1. These steps are listed in table 4. First of all, four steps applied for each word in final list. First step examines whether the word or its lemma exist in NRC or not. If yes, the emotion vector in NRC is considered as work emotion vector. For example, word "Bad" exist in NRC and its emotion vector in NRC is [1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0]. Then, second step examines that word exists in which emotion group of intensity dictionary, and if word exist in any group, its intensity for each emotion group or its corresponding emotion in emotion vector is summed. For example, word "Bad" in intensity dictionary in Anger group has intensity 0.453,

TABLE 4. Extracting the emotion vector for the text in example 1.

Words/steps	Bad		Ugly		Hot		Delicious	
Step1:check word in NRC	EV =[1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0]		EV=[0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]		EV=[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]		EV=[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0]	
Step2: check word in affect list	sense	Affect value	sense	Affect value	sense	Affect value	sense	Affect value
	Anger	0.453	Anger	-	Anger	0.529	Anger	-
	Fear	0.375	Fear	-	Fear	-	Fear	-
	Sadness	0.422	Sadness	-	Sadness	-	Sadness	-
	joy	-	joy	-	joy	-	Joy	0.579
	EV=[1.453, 0 ,1, 1.375, 0 ,1.422, 0, 0, 0, 0, 0, 0]		EV=[0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]		EV=[1. 529, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]		EV=[0, 0, 0, 0, 1.579, 0, 0, 0, 0, 0, 0, 0]	
Step3: check word for affecting by intensifier or reducer	-Do nothing EV=[1.453, 0 ,1, 1.375, 0 ,1.422, 0, 0, 0, 0, 0, 0]		-Do nothing EV=[0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]		Reduce nonzero emotions by 0.2 EV=[1. 329, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]		Intensify nonzero emotions by 0.2 EV=[0, 0, 0, 0, 1.779, 0, 0, 0, 0, 0, 0, 0]	
Step4: EV Final word	EV=[1.453, 0 ,1, 1.375, 0 ,1.422, 0, 0, 0, 0, 0, 0]		EV=[0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]		EV=[1. 329, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]		EV=[0, 0, 0, 0, 1.779, 0, 0, 0, 0, 0, 0, 0]	
Step 5: EV sentence	EV sentence = [2.782 ,0 ,2, 1.375 ,1.779, 1.422, 0, 0 ,0 ,0, 0, 0]							

in Fear group 0.375 and in Sadness group 0.422. Therefore, Anger, Fear and Sadness values from emotion vector or summed intensity values and resulted emotion vector is EV = [1.453, 0, 1, 1.375, 0, 1.422, 0, 0, 0, 0, 0, 0].

In the third step, if the word is influenced by intensifier, all non-zero values in emotion vector are added by 0.2 and if word is influenced by reducer, all non-zero values subtracted by 0.2. Since word "Bad" are influenced by none, in this step, emotion vector does not change. This is while word "Hot" is influenced by reducer and in third step, 0.2 is subtracted from all non-zero values in emotion vector of this word. Also, in third step of emotion vector extraction, non-zero values of word "Delicious" emotion vector that is influenced by intensifier, are added by 0.2. At last, after extracting emotion vector for all words of final list, sentence emotion vector is obtained by summing these vectors that is: EVsentence = [2.782, 0, 2, 1.375, 1.779, 1.422, 0, 0, 0, 0, 0, 0].

The emotional vector is not extracted for the word "uses" because neither the word nor its lemma are present in the NRC, and therefore it will not have any effect even though the word has been in the scope of a reducer. In fact, "uses" is affected by qualifier but it is not in NRC.

After performing this process for all reviews in the dataset, each sentiment column is normalized using (1) to have all values between zero and one.

$$\text{new value} = \frac{\text{old value-min value}}{\text{max value-min value}} \tag{1}$$

3) EXTRACTING VAD VECTOR

For each word in the final list, the *VAD* extraction algorithm is executed

Final list: ['bad', 'ugly', 'hot', 'delicious', 'uses']

it is determined that the word "bad" in the VAD dictionary has three values of [0.125, 0.625, 0.373].

The word ugly in the VAD dictionary has three values of [0.167, 0.63, 0.254].

The word "hot" in the VAD dictionary has three values of [0.49, 0.74, 0.573].

The word delicious in the VAD dictionary has three values of [0.927, 0.65, 0.589]

The VAD vector is not extracted for the word "uses" because it has no value in the VAD dictionary.

Then, by averaging over the VAD vectors of the sentence words, the final VAD vector of the sentence is obtained as follows.

$$\text{VAD\_s}_i : [0.427, 0.661, 0.447]$$

E. EXTRACTING TITLE EMOTIONS

Having an appropriate title can help the review to be read completely. This can have a positive effect on the usefulness of the review. Therefore, in this study, considering the existence of this data in the original dataset, we considered the emotion conveyed by the comment title as a feature. To extract this feature, the same emotion extraction algorithm was used except that, instead of the review text, the input is the review title. Then, since the title is usually short and contains a few words, the emotion vector is sparse (i.e., most of its elements have a value of zero). Therefore, the corresponding emotion of the largest value in this vector is chosen as the title's emotion. If all the emotions have a value of zero, "no sense" will be assigned to the feature.

F. D-S FUSION METHOD FOR PREDICTING THE USEFULNESS OF REVIEWS

Predicting the usefulness of reviews using the D-S score fusion method has the following steps.

- Definition of evidence: At this stage, according to the output of different classification algorithms, evidence is extracted. This is used as basic knowledge in finding the probability of belonging to each class for each review.
- Definition of mass function: According to (2), a mass function is a basic probability assignment for all subsets A of θ [15], [57].

$$\begin{aligned} &m(\mathrm{A}) : 2^{\theta} \rightarrow [0, 1], \mathrm{m}(\phi) = 0, \sum \mathrm{m}(\mathrm{A}) = 1 \\ &2^{\theta} = \{\phi, \{\mathrm{q}_1\} \ldots, \{\mathrm{q}_\mathrm{n}\}, \{\mathrm{q}_1, \mathrm{q}_2\}, \ldots, \\ &\{\mathrm{q}_1, \mathrm{q}_\mathrm{n}\}, \ldots, \{\mathrm{q}_{\mathrm{n}\text{-}1}, \mathrm{q}_\mathrm{n}\}, \{\mathrm{q}_1, \mathrm{q}_2, \mathrm{q}_3\}, \ldots, \\ &\{\mathrm{q}_1, \mathrm{q}_2, \ldots, \mathrm{q}_\mathrm{n}\}\} \end{aligned} \tag{2}$$

D-S theory of evidence is presented by a definite set of mutually exclusive probabilities θ called the detection framework. A subset A of the detection framework θ is called a focal point. 2^{θ} contains all possible subsets of the detection framework θ. A is a component of 2^{θ} and m(A) is the measure of confidence for hypothesis A. m(A) = 0 means that the existing evidence does not support any element of the domain in question while, m(A)= 1 states that the existing evidence only supports A in the domain of interest [57]. In predicting the usefulness of comments, θ is the probability of belonging to one of the five classes.

- Score fusion: we obtain L evidence in the output of the classifiers which are then fused using (3) and (4) [51], [57].

$$m(A) = m_i(B) \oplus m_j(C) = \begin{cases} \dfrac{\Sigma_{B\cap C=A}\, m_i(B)m_j(C)}{1-k} & A \neq \phi \\ 0 & A = \phi \end{cases} \tag{3}$$

$$k = \sum_{B\cap C=\phi} m_i(B)m_j(C) \tag{4}$$

k_{ij} is conflict factor showing the degree of inconsistency between the evidence i and j and is a number between 0 and 1. $k_{ij} = 0$ means that the evidence i and j have no conflicts while, $k_{ij} = 1$ or $0 < k_{ij} < 1$ indicates that two evidence have complete or partial collision to support a review.

As an example for illustrating how the D-S fusion method is applied in review helpfulness prediction settings consider the following review.

"Color Confidence is subtitled "the digital photographer's guide to color management," and is a good overview of the subject. If you want to buy only one book, then Colour Confidence is a good choice. If you want lots of detail, then you're better off buying three separate books - Real World Color Management (Bruce Fraser et al), Professional Photoshop (Dan Margulis, on the subject of colour correction, which Tim Grey only touches on), and Mastering Digital Printing (Harald Johnson)"

The probabilities resulting from the output of the three classifiers based on evidence are shown in Table 5.

TABLE 5. The probabilities of belonging to the 5 classes.

	1	2	3	4	5
S_1	$M_1(1)$ = 0.118	$M_1(2)$ = 0.317	$M_1(3)$ = 0.164	$M_1(4)$ = 0.382	$M_1(5)$ = 0.019
S_2	$M_1(1)$ = 0.168	$M_1(2)$ = 0.205	$M_1(3)$ = 0.197	$M_1(4)$ = 0..186	$M_1(5)$ = 0.244
S_3	$M_2(1)$ = 0.187	$M_2(2)$ = 0.278	$M_2(3)$ = 0.197	$M_2(4)$ = 0.242	$M_2(5)$ = 0.096

S_i is ith evidence.

There are five classes in the table above and the objective is to use the D-S combination rule to get the probability of text belonging to different classes. Therefore, according to (4), the conflict factor is calculated as:

$$\begin{aligned} \mathrm{K} = {} & 0.118 * 0.205 + 0.118 * 0.197 \\ & + 0.118 * 0.186 + 0.118 * 0.244 + 0.317 * \\ & 0.168 + 0.317 * 0.197 + 0.317 * 0.186 \\ & + 0.317 * 0.244 + 0.164 * 0.168 \\ & + 0.164 * 205 + 0.164 * 0.186 + 0.164 * 0.244 \\ & + 0.382 * 0.168 + 0.382 * 0.205 + 0.382 * 0.197 \\ & + 0.382 * 0.244 + 0.019 * 0.168 + 0.019 * 0.205 \\ & + 0.019 * 0.197 + 0.019 * 0.186 = 0.807195 \end{aligned}$$

Also s1 and s2 evidences are fused according to (3).

$$\begin{aligned} \mathrm{m}(1) &= \frac{(0.118 * 0.168)}{(1 - 0.807195)} = 0.1028 \\ \mathrm{m}(2) &= \frac{(0.317 * 0.205)}{(1 - 0.807195)} = 0.3370 \\ m(3) &= \frac{(0.164 * 0.197)}{(1 - 0.807195)} = 0.1675 \\ \mathrm{m}(4) &= \frac{(0.382 * 0.186)}{(1 - 0.807195)} = 0.3685 \\ \mathrm{m}(5) &= \frac{(0.019 * 0.244)}{(1 - 0.807195)} = 0.0240 \end{aligned}$$

Next, the results of the fusion of s1 and s2 evidences are combined with s3 evidence.

$$\begin{aligned}
K = {} & 0.1028*0.278+0.1028*0.197\\
& +0.1028*0.242+0.1028*0.096\\
& +0.3370*0.187+0.3370*0.197\\
& +0.3370*0.242+0.3370*0.096\\
& +0.1675*0.187+0.1675*0.278\\
& +0.1675*0.242+0.1675*0.096\\
& +0.3685*0.187+0.3685*0.278\\
& +0.3685*0.197+0.3685*0.096\\
& +0.0240*0.187+0.0240*0.278\\
& +0.0240*0.197+0.0240*0.242=0.75958
\end{aligned}$$

$$\mathrm{m}(1)=\frac{(0.1028*0.187)}{(1-0.75958)}=0.07$$

$$\mathrm{m}(2)=\frac{(0.3370*0.278)}{(1-0.75958)}=0.38$$

$$m(3)=\frac{(0.1675*0.197)}{(1-0.75958)}=0.13$$

$$m(4)=\frac{(0.3685*0.242)}{(1-0.75958)}=0.37$$

$$m(5)=\frac{(0.0240*0.096)}{(1-0.75958)}=0.0095$$

Therefore, the probability of belonging to classes 2 and 4 is greater for the existing evidence than the other two classes. This is arguable given the initial probabilities for classes 2 and 4. It is clear that the likelihood of scores 1, 3, and 5 decreased after the fusion.

G. TRIPLET STRUCTURE [51]

D-S fusion rule may produce contradictory results in cases there are contradictory evidence. Contradictory evidence in review helpfulness prediction problem may occur when one classifier predicts the helpfulness score of a review to be close to zero (i.e., one star) while other classifiers predicts the score to be completely close to one (i.e., five star). To address this problem and to increase the accuracy of the review usefulness prediction, the triplet structure is used in this study to fuse the evidence. This structure employs the second best decision in combining the classifiers. The improved D-S method using the triple structure is given below [51]:

Definition: If $\{\theta\}$ and $\{\beta\}$ are focal elements and C is the framework of recognition and m is the mass function, the expression in the form $\mathrm{Y}=<\{\theta\},\{\beta\},C>$ is a triplet defined as:

$$\mathrm{m}(\{\theta\})+\mathrm{m}(\{\beta\})+\mathrm{C}=1 \tag{5}$$

The mass function m is called a triplet mass function [51].

*Definition:*If C is the detection framework and we have $|\mathrm{n}|\geq 2$:

$$\varphi_{\mathrm{i}}(\mathrm{d})=\{\mathrm{m}(\{\alpha_1\}),\mathrm{m}(\{\alpha_2\}),\ldots,(\{\alpha_{\mathrm{n}}\})\} \tag{6}$$

In this case, according to the following equation, φ (d) is broken by the law m^{σ}:

$$\{\theta\}=\mathrm{argmax}\ \mathrm{m}(\{\mathrm{a}_1\},(\{\mathrm{a}_2\}),\ldots,(\{\mathrm{a}_{\mathrm{n}}\})\}), \tag{7}$$

$$\{\beta\}=\mathrm{argmax}\ \mathrm{m}(\{\mathrm{a}\}|\mathrm{a}\in\mathrm{a}_1,\ldots,\mathrm{a}_{\mathrm{n}}\}-\{\theta\}) \tag{8}$$

$$m^{\sigma}(\{\theta\})+\mathrm{m}^{\sigma}(\{\beta\})+\mathrm{m}^{\sigma}(\mathrm{C})=1 \tag{9}$$

In particular, m^{σ} is a ternary mass function and is also known as a two-point mass function [51]. So, we have:

$$\varphi_{\mathrm{i}}(\mathrm{d})=\{\mathrm{m}^{\sigma}(\{\theta\}),\mathrm{m}^{\sigma}(\{\beta\}),\mathrm{m}^{\sigma}(\mathrm{C})\},\quad 1\ll\mathrm{i}\ll\mathrm{M} \tag{10}$$

written for the sake of simplicity as:

$$\varphi_{\mathrm{i}}(\mathrm{d})=\{\mathrm{m}(\{\theta\}),\mathrm{m}(\{\beta\}),\mathrm{m}(\mathrm{C})\} \tag{11}$$

To improve the equation for combining two triplet mass functions, we need to consider the relationship between the two single pairs in both triplets. For example, if we have the following two triplets (with the corresponding triplet mass functions m_1 and m_2):

$$<\{\mathrm{a}_1\},\{\mathrm{y}_1\},\mathrm{C}>,<\{\mathrm{a}_2\},\{\mathrm{y}_2\},\mathrm{C}>$$

In this case, the relationship between the two focal pairs $\{\mathrm{a}_1\}$, $\{\mathrm{y}_1\}$, $\{\mathrm{a}_2\}$, and $\{\mathrm{y}_2\}$ is as follows:

1-Two focal points equal:

- If $\{a_1\}=\{a_2\}$ and $\{y_1\}=\{y_2\}$ then $\{a_1\}\cap\{y_2\}=0$ and $\{a_2\}\cap\{y_1\}=0$
- If $\{a_1\}=\{y_2\}$ and $\{y_1\}=\{a_2\}$ then, $\{a_1\}\cap\{a_2\}=0$ and $\{y_2\}\cap\{y_1\}=0$

In this case, the fusion of the two triplet functions consists of three different focal elements. The focal elements $\{\mathrm{a}_1\}$ and $\{\mathrm{y}_1\}$ in one triplet are equal to $\{\mathrm{a}_2\}$ and $\{\mathrm{y}_2\}$ in the other triplet. Two functions of triplet mass m_1 and m_2 are as follows:

$$\mathrm{m}_1(\{\alpha\})+\mathrm{m}_1(\{\mathrm{y}\})+\mathrm{m}_1(\mathrm{C})=1 \tag{12}$$

$$\mathrm{m}_2(\{\alpha\})+\mathrm{m}_2(\{\mathrm{y}\})+\mathrm{m}_2(\mathrm{C})=1 \tag{13}$$

According to Equation (14-17) to combine two triplet masses, we have [51]:

$$\begin{aligned}(\mathrm{m}_1\oplus\mathrm{m}_2)(\{\alpha\}) = {} & \mathrm{K}[\mathrm{m}_1(\{\alpha\})\mathrm{m}_2(\{\alpha\})\\ & +\mathrm{m}_1(\{\alpha\})\mathrm{m}_2(\mathrm{C})+\mathrm{m}_1(\mathrm{C})\ \mathrm{m}_2(\{\alpha\})\end{aligned} \tag{14}$$

$$\begin{aligned}(\mathrm{m}_1\oplus\mathrm{m}_2)(\{\mathrm{y}\}) = {} & \mathrm{K}[\mathrm{m}_1(\{\mathrm{y}\})\mathrm{m}_2(\{\mathrm{y}\})\\ & +\mathrm{m}_1(\{\mathrm{y}\})\mathrm{m}_2\ (\mathrm{C})+\mathrm{m}_1(\mathrm{C})\ \mathrm{m}_2(\{\mathrm{y}\})\end{aligned} \tag{15}$$

$$(\mathrm{m}_1\oplus\mathrm{m}_2)(C)=\mathrm{k}[\mathrm{m}_1(\mathrm{C})\ \mathrm{m}_2(\mathrm{C}) \tag{16}$$

$$\begin{aligned}\mathrm{k}^{-1} = {} & 1-\sum\nolimits_{\mathrm{x}\cap\mathrm{y}=\phi}\mathrm{m}_1(a)\mathrm{m}_2(\mathrm{Y})=1\\ & -\mathrm{m}_1(\{a\})\mathrm{m}_2(\{\mathrm{y}\})\mathrm{m}_2(\{a\})\end{aligned} \tag{17}$$

2- Only one equal focal point:

- If $\{a_1\}=\{a_2\}$ and $\{y_1\}\neq\{y_2\}$ then $\{a_1\}\cap\{y_2\}=0$ and $\{a_2\}\cap\{y_1\}=0$ and $\{y_2\}\cap\{y_1\}=0$
- If $\{a_1\}\neq\{a_2\}$ and $\{y_1\}=\{y_2\}$ then $\{a_1\}\cap\{y_2\}=0$ and $\{a_2\}\cap\{y_1\}=0$ and $\{a_2\}\cap\{a_1\}=0$

- If $\{y_1\} \neq \{a_2\}$ and $\{a_1\} = \{y_2\}$ then $\{y_1\} \cap \{y_2\} = 0$ and $\{a_2\} \cap \{y_1\} = 0$ and $\{a_2\} \cap \{a_1\} = 0$
- If $\{a_1\} \neq \{y_2\}$ and $\{y_1\} = \{a_2\}$ then $\{a_1\} \cap \{y_2\} = 0$ and $\{y_2\} \cap \{y_1\} = 0$ and $\{a_2\} \cap \{a_1\} = 0$

In this case, the combination of two triplet functions consists of four different focal elements. Consider two triplet mass functions m_1 and m_2 with two pairs of $\{a\}$, $\{y\}$ and $\{a\}$, $\{d\}$, $(y \neq d)$. In this case, one focal point in a triplet is equal to another in the other triplet. A general formula for computing the combination of two triplet mass functions is:

$$\begin{aligned}(\mathrm{m}_1 \oplus \mathrm{m}_2)(\{a\}) &= \mathrm{K}[\mathrm{m}_1(\{a\})\mathrm{m}_2(\{a\})\\ &\quad + \mathrm{m}_1(\{a\})m_2(C) + m_1(C)m_2(\{a\})\end{aligned} \tag{18}$$

$$(\mathrm{m}_1 \oplus \mathrm{m}_2)(\{\mathrm{y}\}) = \mathrm{km}_1(\{\mathrm{y}\})\mathrm{m}_2(\mathrm{C}) \tag{19}$$

$$(\mathrm{m}_1 \oplus \mathrm{m}_2)(\{\mathrm{d}\}) = \mathrm{km}_1(\mathrm{C})\ \mathrm{m}_2(\{\mathrm{d}\}) \tag{20}$$

$$(\mathrm{m}_1 \oplus \mathrm{m}_2)(\mathrm{C}) = \mathrm{km}_1(\mathrm{C})\mathrm{m}_2 \tag{21}$$

$$\begin{aligned}\mathrm{K}^{-1} &= 1 - \mathrm{m}_1(\{a\})\mathrm{m}_2(\{\mathrm{d}\}) - \mathrm{m}_1(\{\mathrm{y}\})\\ &\quad \mathrm{m}_2(\{\mathrm{d}\}) - \mathrm{m}_1(\{\mathrm{y}\})\mathrm{m}_2(\{a\})\end{aligned} \tag{22}$$

3- Quite different focal points

If $\{a_1\} \neq \{a_2\}$ and $\{y_1\} \neq \{y_2\}$ and $\{a_1\} \neq \{y_2\}$ and $\{y_1\} \neq \{a_2\}$ then $\{a_1\} \cap \{y_2\} = 0$ and $\{a_2\} \cap \{y_1\} = 0$ and $\{y_2\} \cap \{y_1\} = 0$ and $\{a_2\} \cap \{a_1\} = 0$. In this case, the fusion of two triplet functions consists of five different focal elements and there is no common focal point in the two triplets. If m_1 and m_2 are two triplet functions and $\{a\}$, $\{\mathrm{y}\}$ and $\{\theta\}$, $\{\beta\}$ are two focal pair pairs, the following relationships will exist:

$$\mathrm{m}_1(\{a\}) + \mathrm{m}_1(\{\mathrm{y}\}) + \mathrm{m}_1(\mathrm{C}) = 1 \tag{23}$$

$$\mathrm{m}_2(\{\theta\}) + \mathrm{m}_2(\{\beta\}) + \mathrm{m}_2(\mathrm{C}) = 1 \tag{24}$$

A general formula for computing the combination of two triplet mass functions is:

$$(\mathrm{m}_1 \oplus \mathrm{m}_2)(\{a\}) = \mathrm{km}_1(\{a\})\mathrm{m}_2(\mathrm{C}) = \mathrm{f}(a), \tag{25}$$

$$(\mathrm{m}_1 \oplus \mathrm{m}_2)(\{\mathrm{y}\}) = \mathrm{km}_1(\{\mathrm{y}\})\mathrm{m}_2(\mathrm{C}){=}\mathrm{f}(\mathrm{y}), \tag{26}$$

$$(\mathrm{m}_1 \oplus \mathrm{m}_2)(\{\theta\}) = \mathrm{km}_1(\mathrm{C})\ \mathrm{m}_2(\{\theta\}) = \mathrm{f}(\theta), \tag{27}$$

$$(\mathrm{m}_1 \oplus \mathrm{m}_2)(\{\beta\}) = \mathrm{km}_1(\mathrm{C})\ \mathrm{m}_2(\{\beta\}) = \mathrm{f}(\beta), \tag{28}$$

$$\begin{aligned}\mathrm{k}^{-1} &= 1 - \sum\nolimits_{\mathrm{x}\cap\mathrm{y}=\phi} \mathrm{m}_1(\mathrm{a})\ \mathrm{m}_2(\mathrm{Y}){=}1\\ &\quad - \mathrm{m}_1(\{a\})\mathrm{m}_2(\{\theta\}) - \mathrm{m}_1(\{a\})\mathrm{m}_2(\{\beta\})\\ &\quad - \mathrm{m}_1(\{\mathrm{y}\})\mathrm{m}_2(\{\theta\}) - \mathrm{m}_1(\{\mathrm{y}\})\mathrm{m}_2(\{\beta\})\end{aligned} \tag{29}$$

A practical example of the application of the improved D-S fusion method with the triplet structure for the problem of predicting review usefulness is shown in Table 4. For example, for the following review, the probabilities of predicting the rating by the three classifiers are shown in the following table.

"There is nothing majorly wrong with this game. The plot is well-developed, the characters are customizable, and the battles are strategic. This is probably primarily subjective, but I just didn't enjoy this game. It's not because there were too many movies–it's because I didn't like the movie. I also didn't like the characters or the villains or for that matter the aesthetics. There were some minor but annoying flaws in the game which further contributed to my displeasure. Which button to press was often counterintuitive, and so I often found myself pressing the wrong button. Also, the game badly needs a journal and/or a destination guide so you know where to go–I once spent one hour doing nothing other than walking around a space ship trying to figure out where to go"

According to Table 6, the first and second records in S_1 are of the fifth and fourth classes, respectively. So, we have:

$$\begin{aligned}\mathrm{m}_1(\{a\}) &= \max_{1,1} = 0.42,\\ \mathrm{m}_1(\{\mathrm{y}\}) &= \max_{1,2} = 0.402,\\ Index_{1,1} &= 5, Index_{1,2} = 4,\\ c_1 &= 1 - (\max_{1,1} + \max_{1,2})\\ &= 1 - (0.42 + 0.402) = 0.178\end{aligned}$$

Also for S_2 we have:

$$\begin{aligned}\mathrm{m}_2(\{\mathrm{d}\}) &= \max_{2,1} = 0.244,\\ \mathrm{m}_2(\{a\}) &= \max_{2,2} = 0.205,\\ Index_{2,1} &= 5, Index_{2,2} = 2,\\ c_2 &= 1 - (\max_{2,1} + \max_{2,2})\\ &= 1 - (0.244 + 0.205) = 0.551\end{aligned}$$

Since the condition of an equal focal point is hold, we use the equations (18-22) to fuse S_1 and S_2. So we have:

$$\begin{aligned}k^{-1} &= 1 - (\max_{1,1} * \max_{2,1})\\ &\quad -(\max_{1,2} * \max_{2.1}) - (\max_{1,2} * \max_{2,2})\\ &= 0.71719, (\mathrm{m}_1 \oplus \mathrm{m}_2)(\{5\})\\ &= (1/\mathrm{k}^{-1}) * (\max_{1,1} * \max_{2,2}\\ &\quad + \max_{1,1} * c_2 + \mathrm{c}_1 * \max_{2,2})\\ &= 0.49356, (\mathrm{m}_1 \oplus \mathrm{m}_2)(\{4\})\\ &= (1/\mathrm{k}^{-1}) * \max_{1,2} * c_2 = 0.308840,\\ (\mathrm{m}_1 \oplus \mathrm{m}_2)(\{2\}) &= (1/\mathrm{k}^{-1}) * \max_{2,1} * c_1 = 0.060557,\\ (\mathrm{m}_1 \oplus \mathrm{m}_2)(\mathrm{C}) &= (1/\mathrm{k}^{-1}) * \mathrm{c}_1 * \mathrm{c2} = 0.136750\end{aligned}$$

Next, the results of S_1 and S_2 are fused with S_3. The first and second records (from the outputs of the previous step) are from the fifth and fourth classes, respectively. So, we have:

$$\begin{aligned}\max_{1,1} &= 0.49356, \max_{1,2} = 0.308840,\\ Index_{1,1} &= 5, Index_{1,2} = 4,\\ c_1 &= 1 - (\max_{1,1} + \max_{1,2})\\ &= 1 - (0.49356 + 0.308840) = 0.1976\end{aligned}$$

TABLE 6. Calculated probabilities for a review using the three classifiers.

	Class 1	Class 2	Class 3	Class 4	Class 5	First max	Second max
S_1	0.031	0.115	0.032	0.402	0.42	0.42	0.402
S_2	0.168	0.205	0.197	0.186	0.244	0.244	0.205
S_3	0.209	0.165	0.063	0.413	0.15	0.413	0.209

Also for S_3 we have:

$$\begin{aligned} \max_{2,1} &= 0.413, \max_{2,2} = 0.209 \\ Index_{2,1} &= 4, Index_{2,2} = 1, \\ c_2 &= 1 - (\max_{2,1} + \max_{2,2}) \\ &= 1 - (0.413 + 0.209) = 0.378 \end{aligned}$$

Now, the condition of an equal focal point is hold. So, according to the equations (18-22) we have:

$$\begin{aligned} k^{-1} &= 1 - (\max_{1,1} * \max_{2,1}) \\ &\quad - (\max_{1,2} * \max_{2.1}) - (\max_{1,2} * \max_{2,2}) \\ &= 0.604153, (m_1 \oplus m_2)(\{5\}) \\ &= 0.5479032, (m_1 \oplus m_2)(\{4\}) \\ &= 0.1932317, \\ (m_1 \oplus m_2)(\{1\}) &= 0.13507968, \\ (m_1 \oplus m_2)(C) &= 0.1236322, \\ \max_{Final} &= 0.5479032, \\ Index_{Final} &= 5 \end{aligned}$$

So, the final score,$Index_{Final}$ which is Class 5 is more likely than the other classes.

H. DESCRIPTION OF THE DATASET

In this study, two datasets namely video_games and books were extracted from amazon.com [64]. For each review, rating (5-point scale: 1 to 5 stars), review content, and review title were collected. In total, the first dataset contains 20,000 reviews on the books and related products and the second dataset contains 20,000 reviews on video_games. The description of datasets is shown in Figures 2 and 3.

I. MODELING AND EVALUATION METHODS

In this research different machine learning methods including: decision tree, SVM, random forest, Bagging, naïve Bayes, j48 and AdaBoost were used to construct the classification model. These models are developed in the Python language using the sklearn library [65]. In the evaluations, 10-fold cross validation method is used to prevent overfitting. The criteria for evaluation of the models are the precision, accuracy, F1-score, recall, and Mean Squared Error (MSE) [15]:

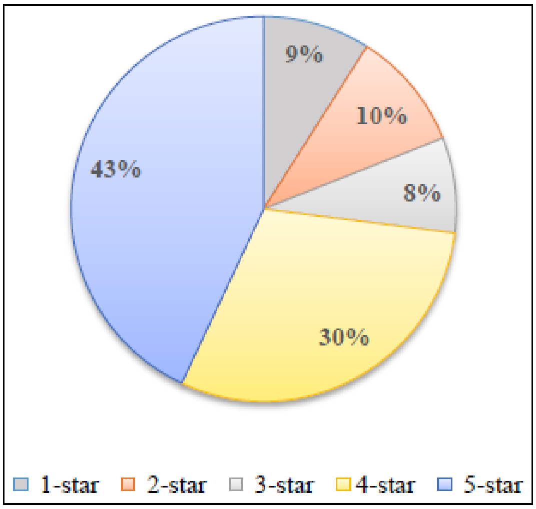

FIGURE 2. Description of the dataset 1 used in this study.

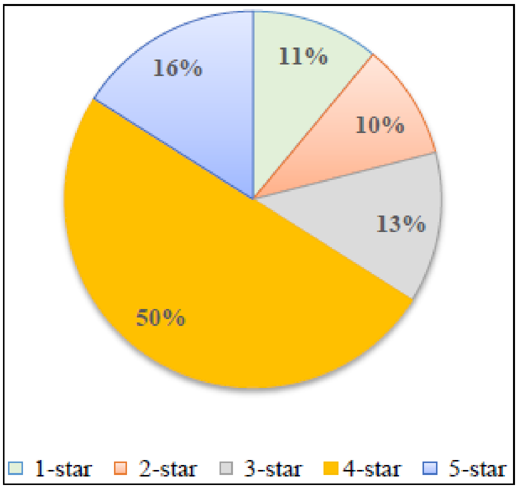

FIGURE 3. Description of the dataset 2 used in this study.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{30}$$

$$Recell = \frac{TP}{TP + FN} \tag{31}$$

$$F1\text{-score} = \frac{2 \times (\Pr ecision \times \operatorname{Re} call)}{\Pr ecision + \operatorname{Re} call} \tag{32}$$

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{33}$$

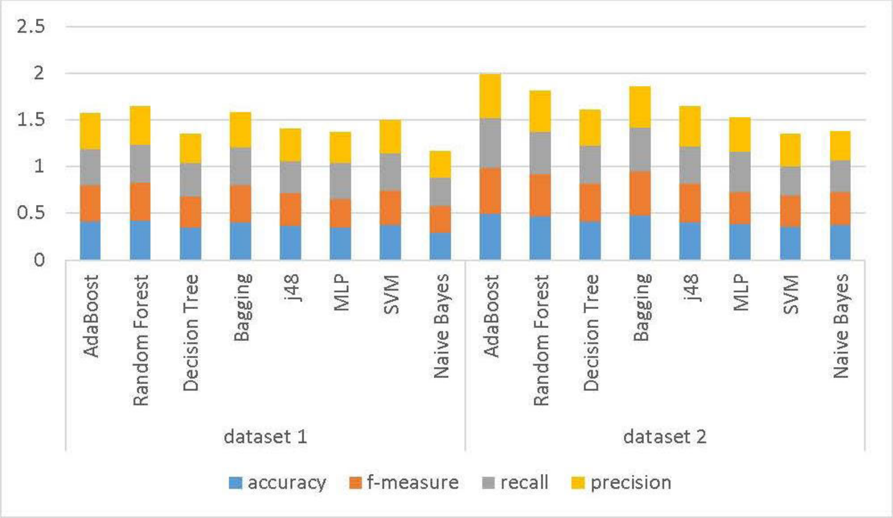

FIGURE 4. Ranking performance by Machine learning algorithms for dataset 1 and dataset 2.

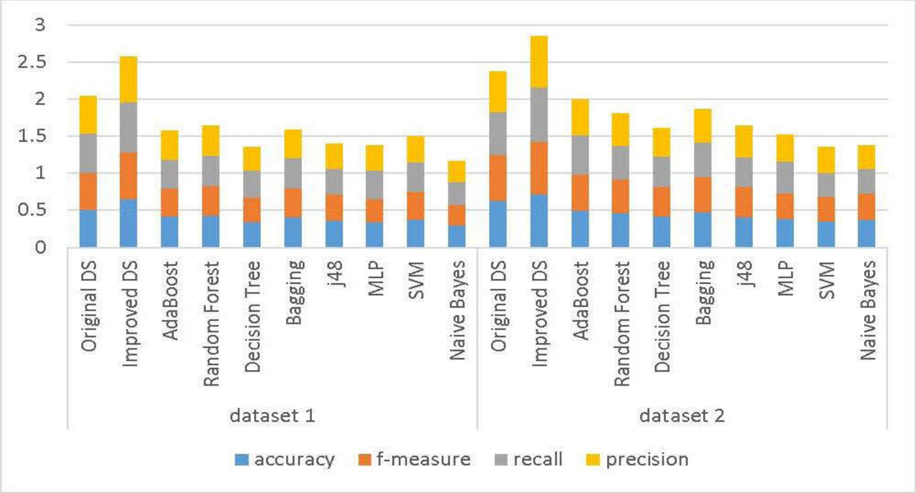

FIGURE 5. Comparison of the results of five-class classification by the Improved D-S method, the original D-S-based method, and Machine learning algorithms on Dataset 1 and Dataset 2.

IV. RESULTS

A. MODELS' PERFORMANCE

The performance of review rating prediction (from 1 to 5 stars) on two datasets using different machine learning methods is shown in Figure 4. These results are compared and the best algorithms is selected to be used later by the improved D-S fusion method. We used all the features in this experiment.

As can be seen, random forest, AdaBoost, and bagging performed best on both datasets. In dataset 1, the accuracy using random forest, Addaboost, and bagging are 0.43, 0.42, and 0.41, respectively. In dataset 2, they are 0.47, 0.5, and 0.48, respectively.

In subsequent evaluations, in order to compare the fusion algorithms with the separate clusters, the reviews are considered in two different scenarios: five-classes and two-classes. Based on the features used, we also created four models for classification: Case1, Case 2, Case 3, Case 4. In Case 1 only text-related features are used, while in Case 2 only VAD and in Case 3 emotion-related features are used. In Case4, all features are employed. In all four modes, text-related features are included.

B. CLASSIFICATION OF 5 CLASSES

Comparison of the results of the 5-class classification using all features (Case 4), by machine learning algorithms, original D-S based and improved D-S method on two-dataset are shown in Figures 5 and 6.

According to Figures 5 and 6, comparing the results of the algorithms shows the superiority of the results of the

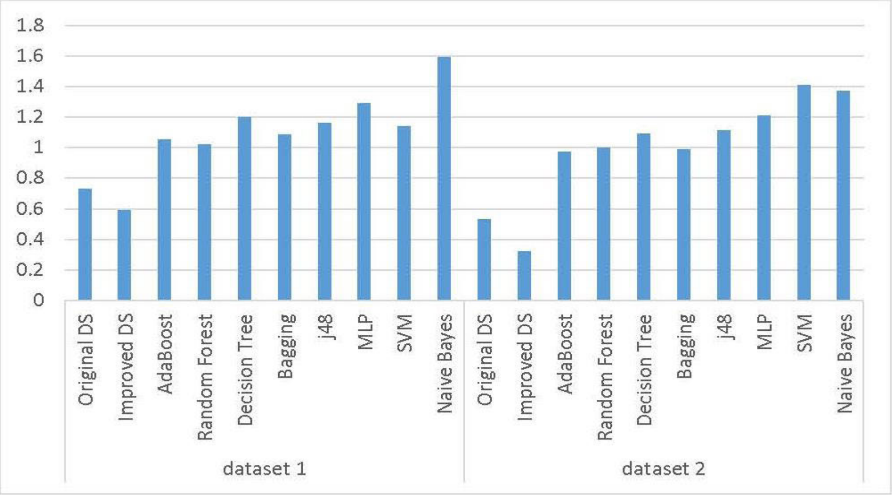

FIGURE 6. Comparison of MSE for improved D-S fusion method and machine learning algorithms (5-class classification) on two datasets.

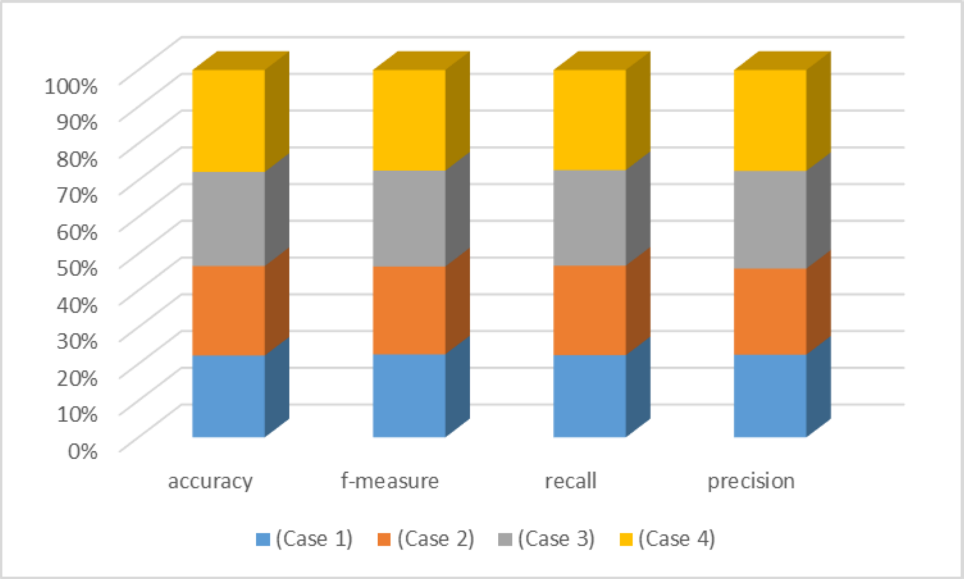

FIGURE 7. Accuracy of the improved D-S fusion method for Cases 1, 2, 3 and 4 Models for Dataset 1.

improved D-S algorithm in ranking the reviews over the original D-S fusion algorithm and machine learning algorithms. As shown in the figures, accuracy, f1-measure, recall and precision on dataset 1 using the improved D-S fusion method are 0.66, 0.63, 0.67, and 0.61, respectively. On dataset 2, these criteria are 0.72, 0.71, 0.73 and 0.69, respectively. Also the MSE criterion on dataset 1 is 0.58 which was improved on dataset 2 by 0.32 using the improved D-S method. Thus, in response to Question 2, it can be said that the improvement of the D-S algorithm using the triplet structure has improved the fusion system and increased the accuracy of the review usefulness prediction system. Hence, we evaluated the other three comparison models only with the improved D-S fusion method.

The test results for each case namely, Case1, Case2, Case3, Case4 are shown in Figures 7 and 8 for dataset 1 and 2, respectively

As shown in Figures 7 and 8, for the first dataset, the accuracy and f-measure criteria were 0.53 and 0.52, respectively, using text-related features. However, in case 2, these criteria are 0.58, 0.55, respectively, and in case 3, using textual and emotion-related features, are 0.61 and 0.6. In case 4 they are 0.66 and 0.63, respectively. Case 4 where all the features were used obtained the best results. Similarly, for the second dataset, in the first case using text-related features, the accuracy and f1-measure criteria were 0.6 and 0.56, respectively. In case 2, these criteria were 0.66 and 0.64 and in case 3, they are 0.69 and 0.66, respectively. Again, the best result obtained

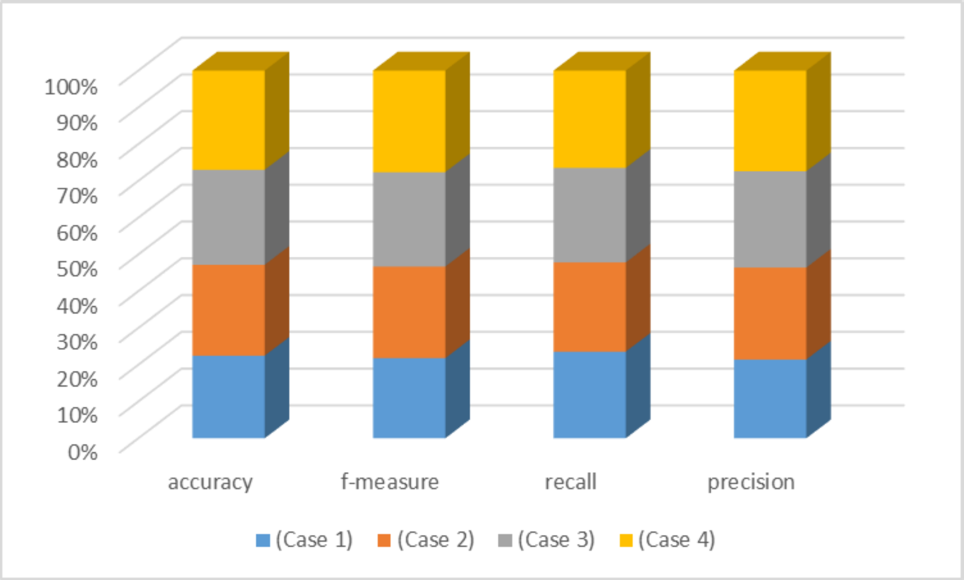

FIGURE 8. Accuracy of the improved D-S fusion method for Cases 1, 2, 3 and 4 Models for Dataset 2.

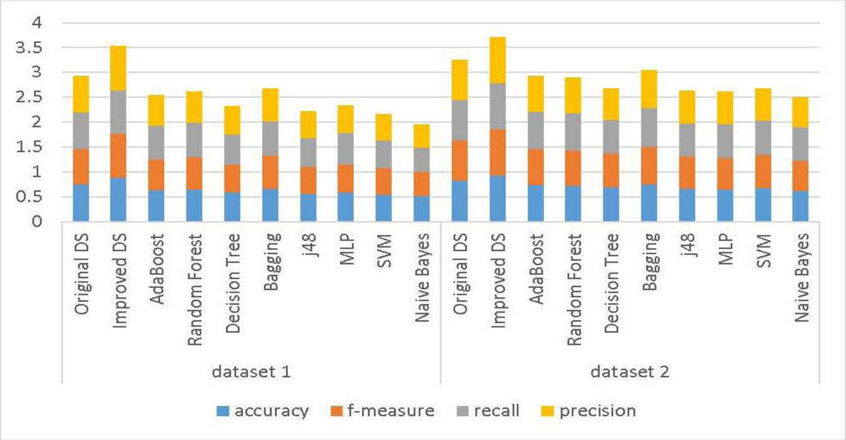

FIGURE 9. Comparison of the results of two-class classification by the Improved D-S method, the original D-S-based method, and machine learning algorithms on Dataset 1 and Dataset 2.

using case 4 where the accuracy and f1-measure criteria were 0.72 and 0.71 (highest score, respectively). Thus, in answer to Question 1, it can be said that considering three dimensions of valence, arousal, and dominance (VAD) along with contextual and emotion-related features affect the usefulness of the review and significantly improve classification accuracy.

C. BINARY CLASSIFICATION

Given that the outputs of this study are 5 classes, classes 2, 1 and 3 are considered as non-useful classes and are shown by 0 and classes 4 and 5 are shown by 1 and are interpreted as useful classes.

Figure 9 shows the results of binary classification based on all properties on two datasets using the original and improved D-S fusion algorithm and separate classifiers.

From the graph, it is clear that the improved D-S fusion algorithm has the best performance in predicting review usefulness on the two datasets and has achieved effective results in improving the fusion system.

As can be seen, the accuracy, f-measure, recall, and precision in the dataset1 using the improved D-S method are 0.89, 0.88, 0.88 and 0.88, respectively. In the dataset2 using the improved D-S method these values are 0.94, 0.93, 0.92 and 0.92, respectively.

Figure 10 illustrates the MSE criterion obtained using the original and improved D-S fusion methods and machine learning algorithms for binary classification. The MSE criterion for books and video_games datasets using the improved D-S method decreased by 15% and 11%, respectively, compared to the original D-S method.

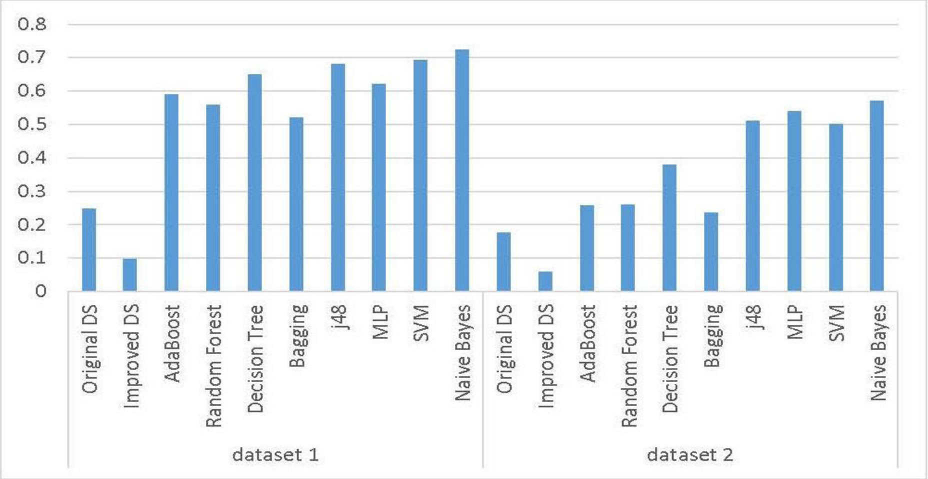

FIGURE 10. Comparison of the MSE of improved D-S fusion method and machine learning algorithms (2-class classification) on two datasets.

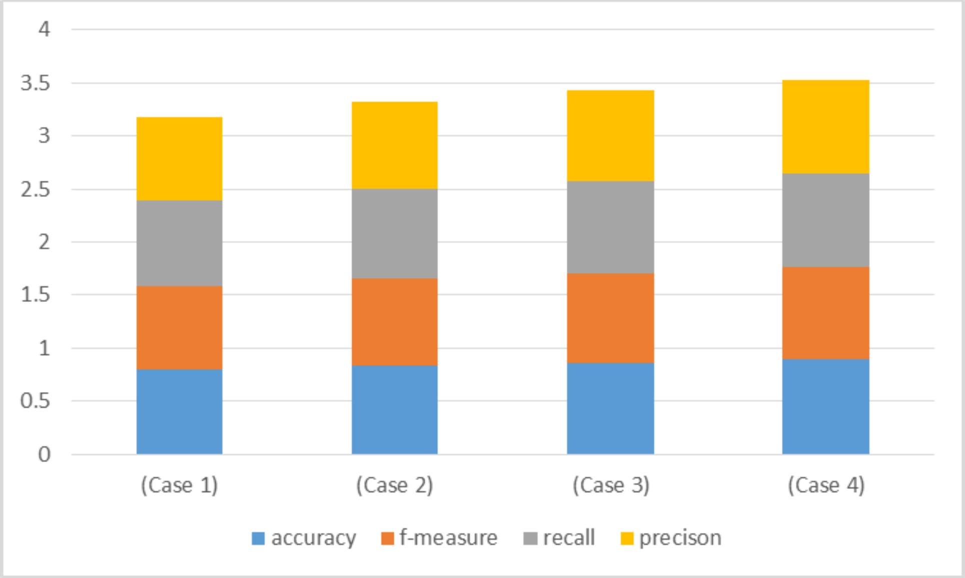

FIGURE 11. Accuracy of the improved D-S fusion method for Cases 1, 2, 3 and 4 for Dataset 1.

The results of the experiments of Case1, Case2, Case3, Case4 using the improved D-S fusion algorithm for dataset1 and dataset2 are shown in Figures 11 and 12.

According to Figures 11 and 12, it is clear that for the first dataset, the accuracy criterion was 0.80 in the first case using the features associated with the review text. However, accuracy in case 2 is 0.83, and in case 3 was increased to 0.86 using contextual and emotion-related features. The best classifier for the model being case 4, where the contextual, VAD, and emotion-related features were used and the accuracy of 0.89 was obtained. Similarly, for the second dataset, the accuracy increased by 0.85 in case 1, 0.89 in case 2, and 0.90 in case 3. The best result was obtained in case 4 where the accuracy is 0.94.

Table 7 shows the feature-wise analysis of 2-class and 5-class classification on two datasets.In this analysis, taking into account the feature or group of features, each time the criteria for the performance of the improved Dempster–Shafer model are obtained, to determine the characteristics or composition of the decisive features.

As shown in table 7, The results of the analysis are similar on the 2-class and 5-class classification on two datasets and it can be seen that the combining the features associated with emotions, features of VAD and text-related features have better helpfulness recognition ability.

V. DISCUSSION

This paper presents a model for predicting review usefulness. To determine review usefulness on emotion-related features such as title's emotion, 12 distinct emotions, and other features such as linguistic features, context-related attributes, valence, arousal, and dominance (VAD) for each review,

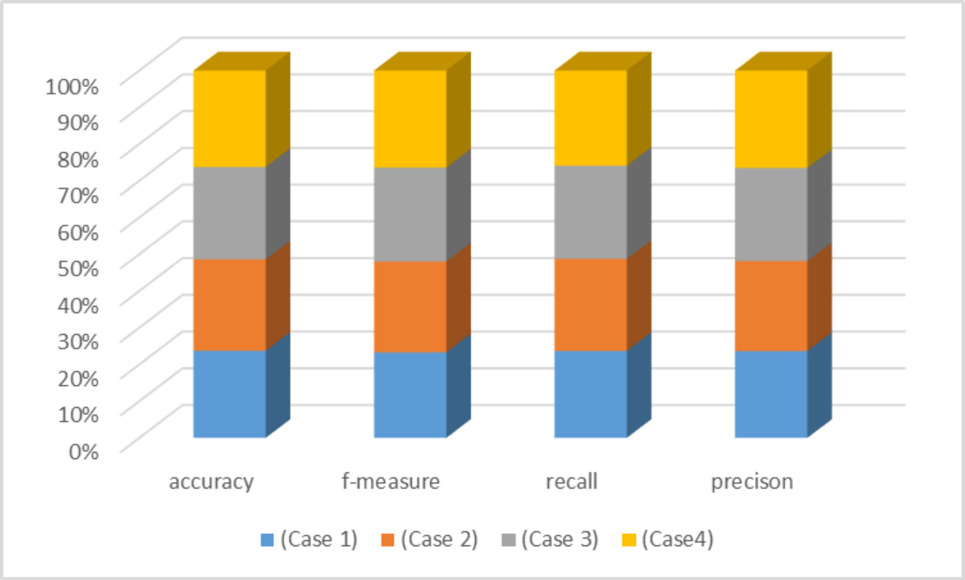

FIGURE 12. Accuracy of the improved D-S fusion method for Cases 1, 2, 3 and 4 for Dataset 2.

length and polarity of opinion is used. Finally, after extracting the required features, predicting review usefulness based on the mentioned features was presented using the improved D-S fusion algorithms and separate classifiers.

We created four models for classification to show the effectiveness of different types of features: Case1, Case2, Case3, Case4, which differ based on the set of properties included as follows; Case 1: Classification using text-related features without including VAD values and emotion-related features. Case 2: Classification with text related features and three VAD dimensions. Case 3: Classification with text related feature and emotion related features. Case 4: Classification using all feature.

The results of the 5-class classification show that the original and improved D-S fusion algorithm and machine learning algorithms have achieved effective results in improving the review helpfulness prediction system. The best result is obtained in case 4, using all features, where on the books and video_games dataset the improved D-S algorithm obtained 15% and 9% higher accuracy than the original D-S algorithm, respectively. The MSE criterion for books and video_games datasets also decreased by 14% and 20%, respectively, compared to the original D-S method.

The best results for 2-class review helpfulness problem were obtained using the improved D-S algorithm with all the features. The MSE criterion for the books and video_games dataset decreased by 15% and 11%, respectively, compared to the baseline method. The accuracy of the classification of books and video_games datasets using the improved D-S algorithm is 14% and 11% higher than the baseline, respectively. Therefore, it can be concluded that these improved results are obtained by exploiting the improved D-S fusion algorithm.

In tables 8, comparing results of 2-class and 5-class classification on two datasets shows that on average, 2-class classification results outperform 5-class classification for both datasets. This may be due to the fact that in 2-class problem, the sensitivity of belonging to different classes is reduced and the likelihood of having an opinion with the predicted class is increased to two existing classes.

Table 9 summarizes the two-class classification results and compares them.

In Table 9, the effect of the proposed features on the two datasets is shown. For the first dataset, accuracy was 0.8 using text-related features. However, the accuracy increased to 0.83 in Case 2 using textual and VAD features and to 0.86 in textual and emotion-related features. The best feature set for the model is Case 4 which used all the features. This Case resulted in accuracy of 0.89. Similarly, for the second dataset, in the first case, accuracy was 0.85 which increased to 0.89 in case 2, and 0.9 in case 3. The best result on this dataset was again obtained in Case 4 where the accuracy reached to 0.94.

The results show that considering the three semantic dimensions of valence, arousal, and dominance (VAD) along with the emotion dimensions and context-related features improves the accuracy of predicting review usefulness scores.

The results were compared with four previous work (Table 6). Ren and Hong [1] used text-related features as well as emotion-related features to predict the usefulness of online consumer opinions and used regression classifiers to classify comments into two categories and reported accuracy of 0.60 and 0.63 on the books and video_games dataset from Amazon. Zhang and Tran worked on text-related features of digital camera reviews from the Amazon site and achieved an accuracy of 0.76 [66]. Ghose and Ipeirotis [33] obtained an accuracy of 0.78 and 0.87 on the DVD, audio, and video and digital camera dataset from the Amazon site. Similary, Krishnamoorthy [35] used linguistic features along with metadata features to predict the usefulness of online

TABLE 7. The importance of features in 2-class and 5-class classification on two datasets.

Feature	Datasets	Measure				Classification
		Precision	Recall	F1-Measure	Accuracy	
Emotion-related features	books	0.57	0.61	0.55	0.57	five-class classification
	video_games	0.64	0.7	0.63	0.66	
	books	0.84	0.86	0.82	0.84	two-class classification
	video_games	0.86	0.89	0.87	0.89	
Text-related features	books	0.5	0.55	0.52	0.53	five-class classification
	video_games	0.54	0.65	0.56	0.6	
	books	0.79	0.81	0.78	0.8	two-class classification
	video_games	0.82	0.84	0.82	0.85	
VAD features	books	0.53	0.58	0.54	0.55	five-class classification
	video_games	0.6	0.65	0.62	0.64	
	books	0.8	0.83	0.8	0.81	two-class classification
	video_games	0.84	0.88	0.84	0.87	
Textual and VAD	books	0.52	0.6	0.55	0.58	five-class classification
	video_games	0.63	0.67	0.64	0.66	
	books	0.82	0.85	0.82	0.83	two-class classification
	video_games	0.85	0.89	0.87	0.89	
Textual and Emotion	books	0.59	0.64	0.6	0.61	five-class classification
	video_games	0.66	0.71	0.66	0.69	
	books	0.86	0.87	0.84	0.86	two-class classification
	video_games	0.88	0.9	0.9	0.9	
all Features	books	0.61	0.67	0.63	0.66	five-class classification
	video_games	0.69	0.73	0.71	0.72	
	books	0.88	0.88	0.88	0.89	two-class classification
	video_games	0.92	0.92	0.93	0.94	

consumer opinions and obtained an accuracy of 0.77 and 0.87 on the Amazon dataset and Blitzer *et al.* [67]. The results show that for both datasets, our method performs better in classifying comments into two categories. Thus, it can be said that the improvement of the D-S algorithm using the triplet structure has improved the fusion system and increased the accuracy of the review usefulness prediction system.

This structure employs the second best decision in combining the classifiers. The benefits of this method are that it not only provides valuable information that is ignored in

TABLE 8. Comparison of the results of two-class classification and five- class classification on two datasets.

Measure	Datasets	Method										Classification
		naive bayse	svm	mlp	j48	bagging	decision tree	random forest	adaBoost	Improved DS	Original DS	
Accuracy	books	0.3	0.38	0.35	0.37	0.41	0.35	0.43	0.42	0.66	0.51	five-class classification
	video_games	0.38	0.36	0.39	0.41	0.48	0.42	0.47	0.5	0.72	0.63	
	books	0.52	0.55	0.6	0.56	0.67	0.59	0.65	0.64	0.89	0.75	two-class classification
	video_games	0.63	0.68	0.65	0.66	0.76	0.69	0.73	0.74	0.94	0.83	
F1-Measure	books	0.28	0.37	0.31	0.35	0.39	0.33	0.4	0.38	0.63	0.5	five-class classification
	video_games	0.35	0.33	0.34	0.41	0.47	0.4	0.45	0.49	0.71	0.62	
	books	0.48	0.53	0.56	0.55	0.66	0.56	0.65	0.62	0.88	0.72	two-class classification
	video_games	0.6	0.66	0.63	0.65	0.75	0.68	0.7	0.73	0.93	0.81	
Recall	books	0.31	0.4	0.38	0.34	0.41	0.36	0.41	0.39	0.67	0.53	five-class classification
	video_games	0.34	0.32	0.43	0.4	0.47	0.41	0.46	0.53	0.73	0.58	
	books	0.5	0.56	0.63	0.58	0.69	0.61	0.69	0.68	0.88	0.73	two-class classification
	video_games	0.67	0.70	0.69	0.67	0.78	0.68	0.76	0.74	0.92	0.82	
Precision	books	0.27	0.35	0.33	0.34	0.37	0.31	0.4	0.38	0.61	0.5	five-class classification
	video_games	0.31	0.34	0.36	0.42	0.44	0.38	0.43	0.47	0.69	0.54	
	books	0.45	0.52	0.55	0.53	0.66	0.56	0.63	0. 6	0.88	0.72	two-class classification
	video_games	0.6	0.63	0.64	0.65	0.75	0.63	0.7	0.72	0.92	0.79	

TABLE 9. Result summarization and comparison.

Approaches	Features			Source	Accuracy
	Textual	VAD*	FE ***		
Two-class classification (Case 1)	✓	×	×	amazon (books)	0.8
Two-class classification (Case 2)	✓	✓	×	amazon (books)	0.83
Two-class classification (Case 3)	✓	×	✓	amazon (books)	0.86
Two-class classification (Case 4)	✓	✓	✓	amazon (books)	0.89
Two-class classification (Case 1)	✓	×	×	amazon (video_games)	0.85
Two-class classification (Case 2)	✓	✓	×	amazon (video_games)	0.89
Two-class classification (Case 3)	✓	×	✓	amazon (video_games)	0.9
Two-class classification (Case 4)	✓	✓	✓	amazon (video_games)	0.94
Ren and Hong [1]	✓	×	✓	amazon (books)	0.60
Ren and Hong [1]	✓	×	✓	amazon (video_games)	0.63
Zhang and Tran [66]	✓	×	×	amazon	0.76
Ghose and Ipeirotis [33]	✓	×	×	amazon	0.78-0.87
Krishnamoorthy [35]	✓	×	×	Blitzer et al. Amazon	0.77-0.87

Note: VAD *: VAD three dimensions; FE **: Feature related features

class labels but also partially avoids the deterioration of performance created by a single prominent class that produces high confidence values.

VI. CONCLUSION

In this study, a model was presented to identify the usefulness of online reviews, using 12 distinct emotions, valence, arousal, and dominance (VAD) vector for each review, other context-related features such as linguistic features, length, and review polarity. Track. Of the 12 mentioned emotions, 8 are from NRC lexicon and 4 are proposed and added in this study as positive surprise, negative surprise, positive expectation, and negative expectation.

In this study, an algorithm was proposed to extract distinct emotions from the text that also improves the emotional intensity of words in different emotion groups. An algorithm for extracting VAD values for each text is also presented. Then, using different machine learning algorithms, the original and improved D-S algorithms different models were developed to predict review helpfulness. Two datasets were used in this study and precision, accuracy, f-score, recall, and Mean Squared Error (MSE) were used to evaluate the results.

According to the results of the five-class and two-class classification, the improved D-S algorithm with triplet structure outperforms the original D-S method and machine learning algorithms. It also improves the accuracy of predicting the usefulness of reviews by combining emotions-related and text-related features.

The overall results for the 2-class scenario is higher than 5-class problem.

Finally, based on obtained results from 5-class and 2-class classification, it could express proposed approach advantages as follow:

- Confirming effectiveness of using word emotion intensity vocabulary in different emotion groups in identifying emotions
- Confirming effectiveness of using VAD vocabulary to extract VAD vector for each text
- Confirming effectiveness of using improved algorithms that consider influential changer on emotions and emotion intensity in different emotion groups in calculations to identifying emotions.
- Confirming effectiveness of using features that related to emotions and VAD in determine review usefulness.
- Increasing precision of review usefulness determine system by improving basic Dempster–Shafer score fusion algorithm.

In future works, we plan to new emotional features introduce and their effect on review usefulness prediction investigate. also, applying deep neural networks to improve emotion recognition system will be investigated as a future work. One of the future works is identification of review usefulness in different types of review and products separately in order to examine more separately and preciously that how much effect different features have on various reviews and products. In addition, applying hybrid evolutional algorithms to increase the accuracy of the review usefulness prediction system is proposed for future research. Other future research is applying proposed method architecture for other languages and also use of proposed variables in other domains such as sentiment analyze, text summarization, recommendation systems and etc.

REFERENCES

[1] G. Ren and T. Hong, "Examining the relationship between specific negative emotions and the perceived helpfulness of online reviews," *Inf. Process. Manage.*, vol. 56, no. 4, pp. 1425–1438, Jul. 2019.

[2] M. E. Basiri, A. R. Naghsh-Nilchi, and N. Ghassem-Aghaee, "A framework for sentiment analysis in persian," *Open Trans. Inf. Process.*, vol. 1, pp. 1–14, Dec. 2014.

[3] Y. Kang and L. Zhou, "Longer is better? A case study of product review helpfulness prediction," Tech. Rep. 2016.

[4] S. Saumya, J. P. Singh, and Y. K. Dwivedi, "Predicting the helpfulness score of online reviews using convolutional neural network," *Soft Comput.*, pp. 1–17, Feb. 2019.

[5] S. Saumya, J. P. Singh, A. M. Baabdullah, N. P. Rana, and Y. K. Dwivedi, "Ranking online consumer reviews," *Electron. Commerce Res. Appl.*, vol. 29, pp. 78–89, May 2018.

[6] M. S. I. Malik and A. Hussain, "Helpfulness of product reviews as a function of discrete positive and negative emotions," *Comput. Hum. Behav.*, vol. 73, pp. 290–302, Aug. 2017.

[7] Y. Hong, J. Lu, J. Yao, Q. Zhu, and G. Zhou, "What reviews are satisfactory: Novel features for automatic helpfulness voting," in *Proc. 35th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. SIGIR*, 2012, pp. 495–504.

[8] J. P. Singh, S. Irani, N. P. Rana, Y. K. Dwivedi, S. Saumya, and P. K. Roy, "Predicting the 'helpfulness' of online consumer reviews," *J. Bus. Res.*, vol. 70, pp. 346–355, Jan. 2017.

[9] Y. Wan, "The matthew effect in social commerce," *Electron. Markets*, vol. 25, no. 4, pp. 313–324, Dec. 2015.

[10] X. Wang, L. R. Tang, and E. Kim, "More than words: Do emotional content and linguistic style matching matter on restaurant review helpfulness?" *Int. J. Hospitality Manage.*, vol. 77, pp. 438–447, Jan. 2019.

[11] M. Lee, M. Jeong, and J. Lee, "Roles of negative emotions in customers' perceived helpfulness of hotel reviews on a user-generated review Website," *Int. J. Contemp. Hospitality Manage.*, vol. 29, no. 2, pp. 762–783, Feb. 2017.

[12] S. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, Jul. 2018, pp. 174–184.

[13] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," *Comput. Intell.*, vol. 29, no. 3, pp. 436–465, Aug. 2013.

[14] S. M. Mohammad and P. D. Turney, "Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon," in *Proc. NAACL HLT Workshop Comput. Approaches Anal. Gener. Emotion Text*, Jun. 2010, pp. 26–34.

[15] M. E. Basiri, A. Kabiri, M. Abdar, W. K. Mashwani, N. Y. Yen, and J. C. Hung, "The effect of aggregation methods on sentiment classification in persian reviews," *Enterprise Inf. Syst.*, pp. 1–28, Oct. 2019.

[16] M. E. Basiri, A. R. Naghsh-Nilchi, and N. Ghasem-Aghaee, "Sentiment prediction based on dempster-shafer theory of evidence," *Math. Problems Eng.*, vol. 2014, pp. 1–13, Apr. 2014.

[17] P. Jamadi Khiabani, M. E. Basiri, and H. Rastegari, "An improved evidence-based aggregation method for sentiment analysis," *J. Inf. Sci.*, vol. 46, no. 3, pp. 340–360, Jun. 2020.

[18] M. E. Basiri, N. Ghasem-Aghaee, and A. R. Naghsh-Nilchi, "Exploiting reviewers' comment histories for sentiment analysis," *J. Inf. Sci.*, vol. 40, no. 3, pp. 313–328, Jun. 2014.

[19] B. Gao, N. Hu, and I. Bose, "Follow the herd or be myself? An analysis of consistency in behavior of reviewers and helpfulness of their reviews," *Decis. Support Syst.*, vol. 95, pp. 1–11, Mar. 2017.

[20] 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," *Electron. Commerce Res. Appl.*, vol. 11, no. 3, pp. 205–217, May 2012.

[21] K. K. Kuan, K.-L. Hui, P. Prasarnphanich, and H.-Y. Lai, "What makes a review voted? An empirical investigation of review voting in online review systems," *J. Assoc. Inf. Syst.*, vol. 16, no. 1, pp. 48–71, 2015.

[22] M. Siering and J. Muntermann, "What drives the helpfulness of online product reviews? From stars to facts and emotions," *Wirtschaftsinformatik*, vol. 7, pp. 103–118, 2013.

[23] D. Weathers, S. D. Swain, and V. Grover, "Can online product reviews be more helpful? Examining characteristics of information content by product type," *Decis. Support Syst.*, vol. 79, pp. 12–23, Nov. 2015.

[24] D. Yin, S. D. Bond, and H. Zhang, "Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews," *MIS Quart.*, vol. 38, no. 2, pp. 539–560, Feb. 2014.

[25] H. Baek, J. Ahn, and Y. Choi, "Helpfulness of online consumer reviews: Readers' objectives and review cues," *Int. J. Electron. Commerce*, vol. 17, no. 2, pp. 99–126, Dec. 2012.

[26] C.-H. Peng, D. Yin, C.-P. Wei, and H. Zhang, "How and when review length and emotional intensity influence review helpfulness: Empirical evidence from Epinions. Com," Tech. Rep., 2014.

[27] Mudambi and Schuff, "Research note: What makes a helpful online review? A study of customer reviews on Amazon.Com," *MIS Quart.*, vol. 34, no. 1, p. 185, 2010.

[28] M. Salehan and D. J. Kim, "Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics," *Decis. Support Syst.*, vol. 81, pp. 30–40, Jan. 2016.

[29] R. M. Schindler and B. Bickart, "Perceived helpfulness of online consumer reviews: The role of message content and style," *J. Consum. Behaviour*, vol. 11, no. 3, pp. 234–243, May 2012.

[30] A. Qazi, K. B. Shah Syed, R. G. Raj, E. Cambria, M. Tahir, and D. Alghazzawi, "A concept-level approach to the analysis of online review helpfulness," *Comput. Hum. Behav.*, vol. 58, pp. 75–81, May 2016.

[31] N. Indurkhya and F. J. Damerau. *Handbook of Natural Language Processing*. Boca Raton, FL, USA: CRC Press, 2010.

[32] A. Ghose and P. G. Ipeirotis, "Designing ranking systems for consumer reviews: The impact of review subjectivity on product sales and review quality," in *Proc. 16th Annu. Workshop Inf. Technol. Syst.*, Dec. 2006, pp. 303–310.

[33] A. Ghose and P. G. Ipeirotis, "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 10, pp. 1498–1512, Oct. 2011.

[34] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou, "Low-quality product review detection in opinion summarization," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn. (EMNLP-CoNLL)*, Jun. 2007, pp. 334–342.

[35] S. Krishnamoorthy, "Linguistic features for review helpfulness prediction," *Expert Syst. Appl.*, vol. 42, no. 7, pp. 3751–3759, May 2015.

[36] Y.-H. Hu and K. Chen, "Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings," *Int. J. Inf. Manage.*, vol. 36, no. 6, pp. 929–944, Dec. 2016.

[37] S. Lee and J. Y. Choeh, "Predicting the helpfulness of online reviews using multilayer perceptron neural networks," *Expert Syst. Appl.*, vol. 41, no. 6, pp. 3041–3046, May 2014.

[38] C. Forman, A. Ghose, and B. Wiesenfeld, "Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets," *Inf. Syst. Res.*, vol. 19, no. 3, pp. 291–313, Sep. 2008.

[39] Y. Wan, B. Ma, and Y. Pan, "Opinion evolution of online consumer reviews in the e-commerce environment," *Electron. Commerce Res.*, vol. 18, no. 2, pp. 291–311, Jun. 2018.

[40] Y. Zhang and Z. Lin, "Predicting the helpfulness of online product reviews: A multilingual approach," *Electron. Commerce Res. Appl.*, vol. 27, pp. 1–10, Jan. 2018.

[41] Y. Pan and J. Q. Zhang, "Born unequal: A study of the helpfulness of user-generated product reviews," *J. Retailing*, vol. 87, no. 4, pp. 598–612, Dec. 2011.

[42] L. M. Willemsen, P. C. Neijens, F. Bronner, and J. A. de Ridder, "'Highly recommended!' the content characteristics and perceived usefulness of online consumer reviews," *J. Comput.-Mediated Commun.*, vol. 17, no. 1, pp. 19–38, Oct. 2011.

[43] Y.-J. Park, "Predicting the helpfulness of online customer reviews across different product types," *Sustainability*, vol. 10, no. 6, p. 1735, 2018.

[44] M. Siering, J. Muntermann, and B. Rajagopalan, "Explaining and predicting online review helpfulness: The role of content and reviewer-related signals," *Decis. Support Syst.*, vol. 108, pp. 1–12, Apr. 2018.

[45] P. Ekman, "An argument for basic emotions," *Cognition Emotion*, vol. 6, nos. 3–4, pp. 169–200, May 1992.

[46] W. G. Parrott. *Emotions in Social Psychology: Essential Readings*. Psychology Press, 2001.

[47] R. Plutchik, *The Emotions: Facts, Theories and a New Model*. New York, NY, USA, 1962.

[48] R. Plutchik. *The Psychology and Biology of Emotion*. New York, NY, USA: HarperCollins College, 1994.

[49] W.-C. Tsao, "Which type of online review is more persuasive? The influence of consumer reviews and critic ratings on moviegoers," *Electron. Commerce Res.*, vol. 14, no. 4, pp. 559–583, Dec. 2014.

[50] E.-J. Lee and S. Y. Shin, "When do consumers buy online product reviews? Effects of review quality, product type, and reviewer's photo," *Comput. Hum. Behav.*, vol. 31, pp. 356–366, Feb. 2014.

[51] Y. Bi, J. Guan, and D. Bell, "The combination of multiple classifiers using an evidential reasoning approach," *Artif. Intell.*, vol. 172, no. 15, pp. 1731–1751, Oct. 2008.

[52] M. E. Basiri, M. Abdar, A. Kabiri, S. Nemati, X. Zhou, F. Allahbakhshi, and N. Y. Yen, "Improving sentiment polarity detection through target identification," *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 1, pp. 113–128, Feb. 2020.

[53] S. Nemati, R. Rohani, M. E. Basiri, M. Abdar, N. Y. Yen, and V. Makarenkov, "A hybrid latent space data fusion method for multimodal emotion recognition," *IEEE Access*, vol. 7, pp. 172948–172964, 2019.

[54] Hatefi, Basiri, and Tamošaitienė, "An evidential model for environmental risk assessment in projects using Dempster–Shafer theory of evidence," *Sustainability*, vol. 11, no. 22, p. 6329, 2019.

[55] D. A. Bell, J. W. Guan, and Y. Bi, "On combining classifier mass functions for text categorization," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 10, pp. 1307–1319, Oct. 2005.

[56] M. Ma and J. An, "Combination of evidence with different weighting factors: A novel probabilistic-based dissimilarity measure approach," *J. Sensors*, vol. 2015, pp. 1–9, Mar. 2015.

[57] J. Yang, H.-Z. Huang, Q. Miao, and R. Sun, "A novel information fusion method based on dempster-shafer evidence theory for conflict resolution," *Intell. Data Anal.*, vol. 15, no. 3, pp. 399–411, May 2011.

[58] S. Nemati and A. R. Naghsh-Nilchi, "Incorporating social media comments in affective video retrieval," *J. Inf. Sci.*, vol. 42, no. 4, pp. 524–538, Aug. 2016.

[59] S. Nemati and A. R. Naghsh-Nilchi, "Exploiting evidential theory in the fusion of textual, audio, and visual modalities for affective music video retrieval," in *Proc. 3rd Int. Conf. Pattern Recognit. Image Anal. (IPRIA)*, Apr. 2017, pp. 222–228.

[60] S. Nemati and A. R. Naghsh-Nilchi, "An evidential data fusion method for affective music video retrieval," *Intell. Data Anal.*, vol. 21, no. 2, pp. 427–441, Mar. 2017.

[61] S. Nemati, "Canonical correlation analysis for data fusion in multimodal emotion recognition," in *Proc. 9th Int. Symp. Telecommun. (IST)*, Dec. 2018, pp. 676–681.

[62] S. M. Mohammad, "Word affect intensities," 2017, *arXiv:1704.08798*. [Online]. Available: http://arxiv.org/abs/1704.08798

[63] U. Farooq, H. Mansoor, A. Nongaillard, Y. Ouzrout, and M. A. Qadir, "Negation handling in sentiment analysis at sentence level," *JCP*, vol. 12, no. 5, pp. 470–478, Sep. 2017.

[64] Amazon.com. *Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs and More*. Accessed: Feb. 10, 2020. [Online]. Available: http://www.amazon.com

[65] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[66] R. Zhang and T. Tran, "Helpful or unhelpful: A linear approach for ranking product reviews," *J. Electron. Commerce Res.*, vol. 11, no. 3, pp. 1–11, Aug. 2010.

[67] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification," in *Proc. 45th Annu. meeting Assoc. Comput. Linguistics*, Jun. 2007, pp. 440–447.

FATEMEH FOULADFAR received the B.S. degree in information technology engineering and the M.Sc. degree in information security from the University of Isfahan, Isfahan, Iran, in 2012 and 2015, respectively. She is currently pursuing the Ph.D. degree in software engineering from the Faculty of Computer Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Iran.

Her current research interests include data mining, data fusion, opinion mining, emotion recognition, and machine learning.

MOHAMMAD NADERI DEHKORDI received the bachelor's and master's degrees in computer engineering, in 1999 and 2001, respectively, and the Ph.D. degree in computer engineering from the Science and Research Branch, Islamic Azad University, Tehran, Iran, in 2009. His Ph.D. Thesis focused on privacy-preserving data mining. He is currently an Assistant Professor and the Dean of the Faculty. He has published over 60 articles in the journal and refereed conference proceedings. He is the author of two research books in mobile database and privacy preserving data mining. His research interests include data mining and knowledge discovery, privacy-preserving data mining/publishing, big data analytics (in the context of scalable, distributed, and mobile platforms), design and configuration of meta-heuristic algorithms in optimization problems, and data analytics, such as novel data mining algorithms, distributed/mobile databases, and social network analysis. He is a Reviewer of some top-ranked journals.

MOHAMMAD EHSAN BASIRI (Member, IEEE) received the B.S. degree in software engineering from Shiraz University, Shiraz, Iran, in 2006, and the M.S. and Ph.D. degrees in artificial intelligence from the University of Isfahan, Isfahan, Iran, in 2009 and 2014, respectively.

Since 2014, he has been an Assistant Professor with the Computer Engineering Department, Shahrekord University, Shahrekord, Iran. He is the author of three books and more than 35 articles. His current research interests include sentiment analysis, natural language processing, machine learning, and data mining.