IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 30, 2020, accepted May 19, 2020, date of publication May 22, 2020, date of current version June 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2996805

Aggregating Customer Review Attributes for Online Reputation Generation

ABDESSAMAD BENLAHBIB AND EL HABIB NFAOUI, (Member, IEEE)
LISAC Laboratory, Department of Computer Science, Faculty of Sciences Dhar EL Mahraz (FSDM), Sidi Mohamed Ben Abdellah University, Fez 30003, Morocco

Corresponding author: Abdessamad Benlahbib (abdessamad.benlahbib@usmba.ac.ma)

ABSTRACT In this paper, we face the problem of generating reputation for movies, products, hotels, restaurants and services by mining customer reviews expressed in natural language. To the best of our knowledge, previous studies on reputation generation for online entities have primarily examined semantic and sentiment orientation of customer reviews, disregarding other useful information that could be extracted from reviews, such as review helpfulness and review time. Therefore, we propose a new approach that combines review helpfulness, review time, review attached rating and review sentiment orientation for the purpose of generating a single reputation value toward various entities. The contribution of the paper is threefold. First, we design two equations to compute review helpfulness and review time scores, and we fine-tune Bidirectional Encoder Representations from Transformers (BERT) model to predict the review sentiment orientation probability. Second, we design a formula to assign a numerical score to each review. Then, we propose a new formula to compute reputation value toward the target entity (movie, product, hotel, restaurant, service, etc). Finally, we propose a new form to visualize reputation that depicts numerical reputation value, opinion categories, top positive review and top negative review. Experimental results coming from several real-world data sets of miscellaneous domains collected from IMDb, TripAdvisor and Amazon websites show the effectiveness of the proposed method in generating and visualizing reputation compared to three state-of-the-art reputation systems.

INDEX TERMS Reputation generation, text mining, sentiment analysis, natural language processing, BERT encoder, decision making, e-commerce.

I. INTRODUCTION

The exponential growth of Web 2.0 has dramatically impacted the evolution of e-commerce platforms [1]–[4]. Recent online shopping statistics showed that the number of users of some famous e-commerce websites such as Jingdong,[1] Alibaba[2] and Amazon[3] has exceeded 1 billion [5]. Thereby, customer reviews attached to a product can easily surpass thousands [2], [6], [7]. In fact, while, a good number of reviews could indeed give a hint about the quality of an item, a potential customer may not have the time or effort to read all reviews for the purpose of making a decision [8]. Thus, the need for the right tools and technologies to help in such a task becomes a necessity for the buyer as for the seller.

Currently, little work has been performed to support customer decision making in E-commerce using natural language processing techniques. We identify principally two techniques. The first one is feature based summarization that aims to identify the target entity (product, movie, hotel, restaurant, service) features and its corresponding opinions polarity (positive/negative), then, a feature-based summary of the reviews is generated [3], [4], [7]. While the second technique is called reputation generation, whose main focus is to produce an estimation value in which an entity is held based on mining customer reviews expressed in natural languages [5], [9]–[11].

Previous studies on reputation generation have primarily focused on using semantic and sentiment analysis [5], [9]–[11], disregarding other useful information that could be extracted from user reviews, such as *"review helpfulness,*

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

[1]https://www.jd.com
[2]https://www.alibaba.com/
[3]https://www.amazon.com

which implies that reviews that receive higher votes from other users typically provide more information", and *"review time, which implies that more recent reviews generally provide users with more up-to-date information"*.

An accurate and reliable reputation system should consider exploiting more online reviews features such as review attached rating, review helpfulness, review time and review sentiment orientation. For that reason, we propose a reputation system that incorporates all these attributes during the process of generating and visualizing reputation for various entities (movies, products, hotels, restaurants and services). In this manner, this study addressed the following research question: with the consideration of review helpfulness, review time, review sentiment orientation probability and review attached rating, can the proposed reputation system offer better results in terms of reputation generation and visualization than the previous reputation systems (consider only semantic and sentiment relations)?

The contributions of this work are summarized as follows:

- Firstly, we propose a novel system that incorporates review time, review helpfulness, review sentiment orientation and review attached rating for the purpose of generating a numerical reputation value toward various entities (movies, products, hotels, restaurants, services, etc).
- Secondly, we propose a new holistic form to visualize reputation by showing numerical reputation value, opinion categories, top positive review and top negative review in order to support customers during their decision making process in E-commerce (buying, renting, booking).

The article is organized in the following way: Section 2 gives a literature review of related work for document level sentiment analysis and natural language processing techniques for decision making in E-commerce. Section 3 presents problem statement. In section 4, we elaborate our reputation system. The conducted experiments and discussion are presented in section 5. In section 6, the conclusion of this work is provided.

II. LITERATURE REVIEW

This section describes and examines previous research work done in the area of natural language processing (NLP) techniques for decision making in E-commerce and document level sentiment analysis.

A. NLP TECHNIQUES FOR DECISION MAKING IN E-COMMERCE

The BusinessDictionary[4] defines decision making as: *"The thought process of selecting a logical choice from the available options"*. During the last twenty years, few approaches have been proposed to help potential customers making decisions in E-commerce websites using mainly two NLP techniques: feature-based summarization of customer reviews and reputation generation.

[4]http://www.businessdictionary.com/definition/decision-making.html

Hu and Liu (2004) [7] were the first to design and build a system that produces a feature-based summary from customer reviews. The proposed system performs three tasks: (1) association rule mining [12] is used to extract product features from customer reviews, (2) WordNet [13] is utilized to predict the semantic orientations of opinion words, (3) a feature-based summary is produced. Over the last two decades, few systems have been proposed to perform feature-based summarization. The summarizers are applied on various domains: product reviews [7], [14]–[17], movie reviews [4], local services reviews [18] and hotel reviews [2], [19], etc.

Backing to reputation generation. The pioneer work that tackles the task of reputation generation based on mining opinions expressed in natural languages was firstly proposed by Yan *et al.* (2017) [5] in which reviews are fused into different opinion sets based on their semantic relations, then, a single reputation value is generated by aggregating the fused and grouped opinions statistics (sum of similarities, sum of ratings, number of reviews). In [9], the authors applied K-means clustering algorithm to group similar reviews into the same cluster using Latent Semantic Analysis (LSA) before producing a reputation value using the statistics of each cluster. However, both approaches have relied on extracting semantic relations between reviews and have disregarded the fact that the majority of online customer and user reviews are opinionated. Benlahbib and Nfaoui (2019) [10] proposed a fourfold approach to improve [5]. First, Naïve Bayes and Linear Support Vector Machines classifiers were applied to separate reviews into positives and negatives by predicting their sentiment polarity. Second, positive and negative reviews were fused into different sets based on their semantic similarity (Latent Semantic Analysis and cosine similarity). Third, a custom reputation value is computed separately for both positive opinion sets and negative opinion sets. Finally, a single reputation value is calculated using the weighted arithmetic mean.

Since all of the above-mentioned reputation systems exploit only semantic and sentiment features, we propose a new reputation system that incorporates more features for the purpose of generating an accurate and reliable reputation value toward various entities.

B. DOCUMENT LEVEL SENTIMENT ANALYSIS

Ahlgren (2016) [20] defines sentiment analysis as: *"the process of identifying and detecting subjective information using natural language processing, text analysis, and computational linguistics"*. Generally, sentiment analysis can be divided into three levels: sentence level opinion mining, document level opinion mining and fine-grained opinion mining. Since we have applied document level sentiment analysis to extract the sentiment orientation of customer and user reviews, this section will mainly focus on previous research work done in the area of document level opinion mining.

According to [21], document level opinion mining is: *"a task of extracting the overall sentiment polarities of given*

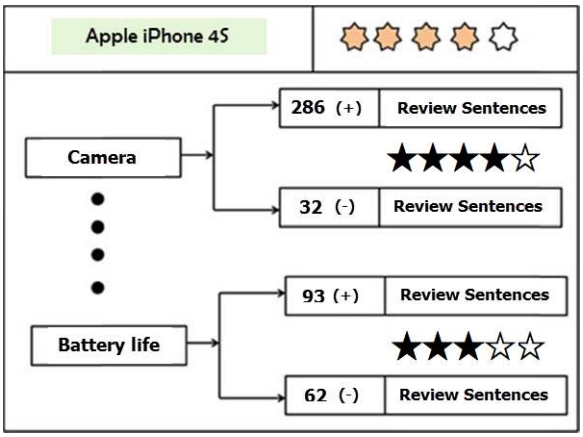

FIGURE 1. Kangale *et al.* (2016) [6] feature-based summary.

documents, such as movie reviews, product reviews, tweets and blogs.".

Many approaches have been used to handle the task of document level sentiment analysis:

- Supervised approaches: These approaches require annotated corpus to train machine learning models. The first work for supervised document level opinion mining was proposed by Pang *et al.* (2002) [22]. Three machine learning classifiers (Support Vector Machines (SVMs) [23], Naïve Bayes classifier [24] and Maximum Entropy classifier [25]) were trained with movie reviews labeled by sentiment (positive/negative). The authors trained the three models on various kinds of features (unigrams, bigrams, parts of speech and position) and found that the sentiment classification task preforms well when adopting unigrams as features. Kennedy and Inkpen (2006) [26] trained Support Vector Machine classifiers on unigrams and bigrams by incorporating three types of context valence shifters: *"intensifiers"*, *"negations"* and *"diminshers"*. The trained model achieved an accuracy of 0.859 on movie review data[5] [27]. Koppel & Schler (2006) [28] defined the sentiment classification task as a three-category problem (positive, negative and neutral) and used different learning algorithms: SVM, J48 Decision Tree [29]. Naïve Bayes, Linear Regression [30] and Frank and Hall (2001) [31] classification method. The results show significant improvement on the sentiment classification accuracy when using neutral examples over ignoring them. In [32], the authors combined Naïve Bayes and Support Vector Machine by training SVM with Naïve Bayes log-count ratios as features. The proposed model has achieved promising results across several datasets. Jing *et al.* (2015) [33] applied Naïve Bayes algorithm on 3046 customer reviews related to fifty-eight business-to-team (B2T) websites to study the survival conditions of B2T companies. Augustyniak *et al.* (2016) [34] presented a wide comparison and analysis of opinion mining task for several classifiers: Random Forests [35], Linear SVC, Bernoulli Naïve Bayes, Mulinomial Naïve Bayes, Extra Tree Classifier [36], Logistic Regression and AdaBoost [37], [38]. They conducted experiment on Amazon reviews dataset [39] and found that Logistic Regression classifier outperforms the other classifiers in predicting sentiment polarity of product reviews.
- Unsupervised approaches: They attempt to determine the sentiment orientation of a text by applying a set of rules and heuristics obtained from language knowledge. Turney (2002) [40] was the first to propose an unsupervised sentiment analysis technique to classify reviews as *"recommended"* or *"not recommended"*. The semantic orientation of a phrase is computed as the pointwise mutual information (PMI) [41] between the given phrase and the word *"excellent"* minus the pointwise mutual information between the given phrase and the word *"poor"*. The proposed algorithm achieved an accuracy of 84% for automobile reviews, 80% for bank reviews, 71% for travel destination reviews and 66% for movie reviews. In [42], the authors proposed a lexicon-based method to opinion mining text by using a dictionary of sentiment words and their semantic orientations varied between -5 and $+5$. The authors also incorporated amplifiers, downtoners and negation words to compute a sentiment score for each document. Vashishtha and Susan (2020) [43] proposed a fuzzy rule-based approach to perform opinion mining of tweet. The authors use a novel unsupervised nine fuzzy rule based system to predict the sentiment orientation of the post (positive, negative or neutral). In [44], Fernández-Gavilanes *et al.* (2016) proposed a sentiment analysis approach to predict the polarity in online textual messages such as tweets and reviews using an unsupervised dependency parsing-based text classification method.
- Deep learning approaches: Over the past few years, deep learning models have greatly improved the state-of-the-art of opinion mining. Moraes *et al.* (2013) [45] made a comparative study between Support Vector Machines (SVM) and Artificial Neural networks (ANN) for document-level opinion mining and found that ANN results are at least comparable or superior to SVMs. To overcome the weakness of bag-of-words (BoW) model, the authors [46] proposed an unsupervised algorithm named paragraph vector (doc2vec), an extension to word2vec approach [47]. The proposed algorithm learns vector representations for variable-length texts such as sentences, paragraphs, and documents. Experimental results depict that doc2vec algorithm achieved new state-of-the-art results on several sentiment analysis tasks. Johnson and Zhang (2015) [48] trained a parallel Convolutional Neural Network (CNN) [49] without using pre-trained word vectors: word2vec, doc2vec and GloVe[6] [50]. Instead, convolutions are

[5]http://www.cs.cornell.edu/people/pabo/movie-review-data/

[6]https://nlp.stanford.edu/projects/glove/

directly applied to one-hot encoding vectors to leave the network solely with information about the word order. The proposed approach achieved an accuracy rate of 92.33% on Large Movie Review Dataset[7] outperforming both SVM [51] and NB-LM [52]. Baktha and Tripathy (2017) [53] investigated the performance of Long Short-Term Memory (LSTM) [54], vanilla RNNs and Gated Recurrent Units (GRU) on the Amazon health product reviews dataset and sentiment analysis benchmark datasets SST-1 and SST-2. The results depict that GRU achieved the highest sentiment classification accuracy. In [55], the authors combined unsupervised data augmentation (UDA) with Bidirectional Encoder Representations from Transformers (BERT) [56] and compared it with fully supervised $BERT_{LARGE}$ on six text classification benchmark datasets. The authors reported that their proposed approach outperforms $BERT_{LARGE}$ on five text classification benchmark datasets including Large Movie Review Dataset. Facebook AI and University of Washington researchers [57] improved BERT by proposing *"Robustly Optimized BERT approach"* (RoBERTa) that was trained with more data and more number of pretraining steps and has dropped the next sentence prediction (NSP) approach used in BERT. Recently, Google researchers [58] proposed XLNet, a *"generalized autoregressive pretraining method"* that outperforms Bidirectional Encoder Representations from Transformers (BERT) on twenty text classification tasks, and achieves state-of-the-art results on eighteen text classification tasks including sentiment analysis. At the end of 2019, Lan *et al.* (2019) [59] proposed ALBERT: *"A Lite BERT for Self-supervised Learning of Language Representations"*. The paper describes parameter reduction techniques to lower memory reduction and increase the training speed and accuracy of BERT models. In [60], the authors introduced a novel *"Text-to-Text Transfer Transformer"* (T5) neural network model pre-trained on a large text corpus which can convert any language problem into a text-to-text format. The T5 model achieved state-of-the-art results on SST-2 Binary classification dataset with an accuracy of 97.4%. Recently, Clark *et al.* (2020) [61] presented *ELECTRA* that uses new pre-training task called replaced token detection (RTD). The experiment results showed that RTD is more efficient than masked language modeling (MLM) pre-training models such as BERT.

III. PRELIMINARIES

This section covers the necessary background for understanding the remainder of the paper, including the problem definition and BERT model which is fine-tuned to determine the sentiment orientation of customer and user reviews in our proposed system.

[7]https://ai.stanford.edu/~amaas/data/sentiment/

A. PROBLEM DEFINITION

In this paper, we face the problem of generating reputation for movies, products, hotels, restaurants and services by aggregating review time, review helpfulness votes, review sentiment orientation and review attached rating. Given a set of reviews $R_j = \{r_{1j}, r_{2j}, \ldots, r_{nj}\}$ expressed for an entity E_j, the set of their attached ratings $V_j = \{v_{1j}, v_{2j}, \ldots, v_{nj}\}$ where $v_{ij} \in [1, 5]$ or $v_{ij} \in [1, 10]$ depending on the rating system, the set of their attached helpfulness votes $RH_j = \{rh_{1j}, rh_{2j}, \ldots, rh_{nj}\}$ where $rh_{ij} \in \mathbb{N}^*$, the set of their posting time $RT_j = \{rt_{1j}, rt_{2j}, \ldots, rt_{nj}\}$ and the set of their sentiment orientation probabilities predicted by fine-tuned $BERT_{base}$ model $BERT_j = \{bert(r_{1j}), bert(r_{2j}), \ldots, bert(r_{nj})\}$ where $bert(r_{ij}) \in [0, 1]^2$. The goal is to compute a review score for each review $RS_j = \{rs_{1j}, rs_{2j}, \ldots, rs_{nj}\}$ based on its helpfulness votes, its posting time and its sentiment orientation, and finally, compute a reputation value *Rep* for an entity *j* by averaging the product of reviews score and reviews attached rating. Table 1 presents the descriptions of notations used in the rest of this paper.

TABLE 1. Symbol denotation.

Symbol	Description
R_j	The set of reviews expressed for the entity j
V_j	The set of ratings expressed for the entity j
RH_j	The set of reviews helpfulness votes expressed for the entity j
RT_j	The set of reviews posting time expressed for the entity j
$BERT_j$	The set of sentiment orientation probabilities predicted by fine_tuned $BERT_{Base}$ for reviews expressed for the entity j
RS_j	The set of reviews score expressed for the entity j
E_j	The target entity j
O_j	The total number of reviews expressed for the target entity j
Rep	The reputation value

B. BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT)

Bidirectional Encoder Representations from Transformers (BERT) is: *"a new language representation model which is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications"* [56].

According to [56]: *"BERT uses a bidirectional Transformer. OpenAI GPT [62] uses a left-to-right Transformer. ELMo [63] uses the concatenation of independently trained left-to-right and right-toleft LSTMs to generate features for downstream tasks. Among the three, only BERT representations are jointly conditioned on both left and right context in all layers. In addition to the architecture differences, BERT and OpenAI GPT are fine-tuning approaches, while ELMo is a feature-based approach"* (Figure 2).

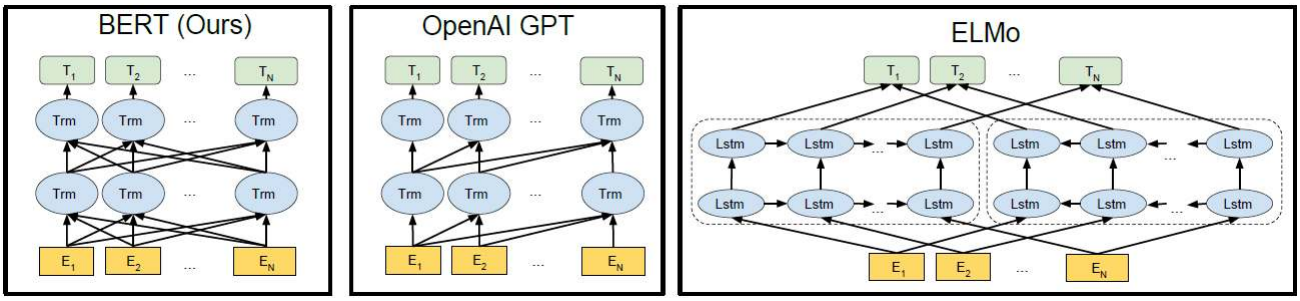

FIGURE 2. Differences in pre-training model architectures [56].

IV. PROPOSED APPROACH

A. SYSTEM OVERVIEW

Our approach consists mainly on four steps:

- Firstly, we collect real data from websites that specialize in gathering customer reviews such as IMDb,[8] TripAdvisor[9] and Amazon[10] using web scraping tools, then, we preprocess them.
- Secondly, we assign three numerical scores to each review: helpfulness score, time score and sentiment orientation score.
- Thirdly, we compute a review score based on the precomputed scores (helpfulness score, time score and sentiment orientation score).
- Finally, we generate a numerical reputation value toward the target entity (product, movie, hotel, restaurant, service, etc). Then, we propose a new form to visualize reputation by depicting numerical reputation value, opinion categories, positive review with the highest score and negative review with the highest score.

Figure 3 describes the pipeline of our work.

B. DATA COLLECTION AND PREPROCESSING

Differently from previous studies on reputation generation, which mainly focus on extracting semantic and sentiment relations of reviews, our work incorporates other factors such as review helpfulness and review time. Hopefully, the majority of popular E-commerce websites such as TripAdvisor[11] and Amazon[12] gather online reviews with respect to the following structure: **textual review, review helpfulness votes and review posting time**. Figure 4 describes online reviews structure.

With the use of a web scraping tool, we have been able to collect raw data from some real data suppliers like Amazon, TripAdvisor and IMDb.

After collecting all reviews, we applied some preprocessing techniques (lowercasing, tokenization, ...). Technical details of data collection and preprocessing phase are described in section V. EXPERIMENT RESULTS subsection A. EXPERIMENTAL DATA COLLECTION AND PREPROCESSING.

[8]https://www.imdb.com/
[9]https://www.tripadvisor.com/
[10]https://www.amazon.com
[11]https://www.tripadvisor.com/
[12]https://www.amazon.com

C. REVIEW HELPFULNESS

The number of helpfulness votes attached to a review indicates how informative it is, which implies that reviews that receive higher votes from other users typically provide more information. Thus, we design formula (1) to compute review helpfulness score.

$$H(r_{ij}) = \begin{cases} 0.75 & \text{if } rh_{ij} = 0 \text{ or } \dfrac{\log_{10}(rh_{ij})}{\log_{10}(N_j)} \leqslant 0.75 \\ \log_{N_j}(rh_{ij}) & \text{Otherwise} \end{cases} \quad (1)$$

We denote:

r_{ij}: Review number i expressed for the entity j.

$H(r_{ij})$: Helpfulness score of review r_{ij}.

rh_{ij}: The number of helpfulness votes attached to review r_{ij}.

N_j: The number of helpfulness votes attached to the most voted review toward the entity j.

The helpfulness score for a review ranges between 0.75 and 1 because we don't want to assign a low score to reviews with a small number of helpfulness votes.

We mention that $\frac{\log_{10}(rh_{ij})}{\log_{10}(N_j)} \leqslant 0.75$ means that $\log_{N_j}(rh_{ij}) \leqslant 0.75$ due to the fact that:

$$\log_N(n) = \frac{\log_{base}(n)}{\log_{base}(N)} = \frac{\log_{10}(n)}{\log_{10}(N)}$$

By applying equation (1), the most voted review will receive a review helpfulness score of 1 since $\log_{N_j}(N_j) = 1$. Reviews with high helpfulness votes will receive a high review helpfulness score and reviews with low helpfulness votes will receive a low review helpfulness score since for $x \in [1, N]$ and $y \in [1, N]$: $x \leq y$ implies that $log_N(x) \leq log_N(y)$.

Algorithm 1 computes the helpfulness score for review r_{ij}.

D. REVIEW TIME

Could you tell what would happen if we take a very well-reviewed gaming laptop from 10 years ago and put it on an online store? To answer this question, let us travel back in time to 20 years ago, where the gaming industry witnessed a great competition between gaming consoles, and where the

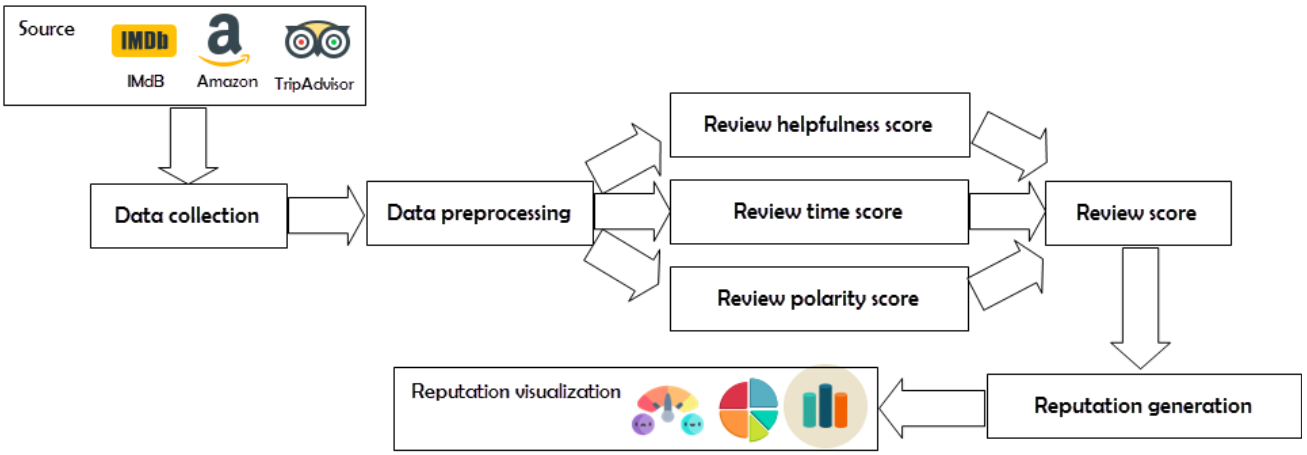

FIGURE 3. Proposed system pipeline.

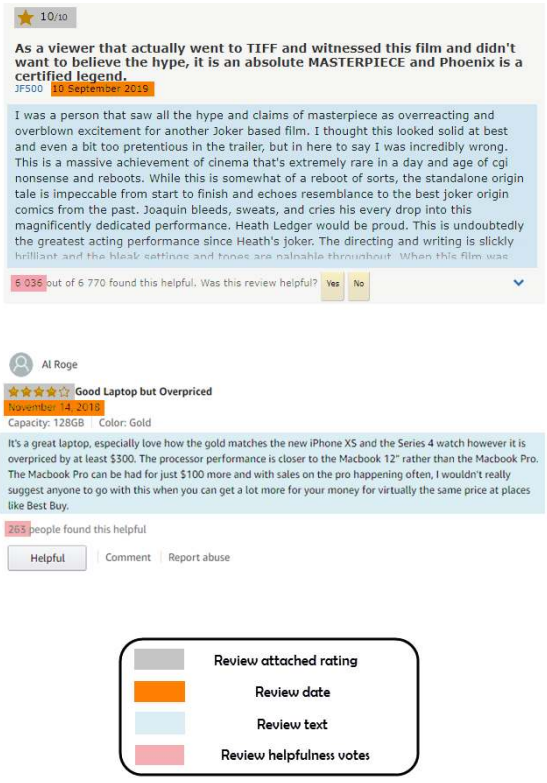

FIGURE 4. IMDb and Amazon reviews structure.

enjoyment of a hardcore gaming experience was limited to that kind of tech. In that era, a gamer had to have a fat and heavy TV, with cables attached to a relatively big dedicated gaming console and its controllers in order to play a video game. All this bunch of materials and cables remain in one place in the house. Next, the industry shifted to computers, and then to mobile computers also known as laptops, which brought enough satisfaction to all the gaming consumers over the world. Although, laptops have been made much heavier than they should be, yet, it was very exciting to have the ability to enjoy your favorite games wherever you want just by packing your laptop on a backpack rather than having to

Algorithm 1 Review Helpfulness Score

Define: $RH_j = \{rh_{1j}, rh_{2j}, \ldots, rh_{nj}\}$: The set of reviews helpfulness votes expressed for the entity j.
Input: RH_j
Function $\mathrm{H}(rh_{ij})$ **:**
  if $rh_{ij} = 0$ *or* $\frac{\log_{10}(rh_{ij})}{\log_{10}(max(RH_j))} \leqslant 0.75$ **then**
    $H \longleftarrow 0.75$
  else
    $H \longleftarrow \log_{max(RH_j)}(rh_{ij})$
  end if
  return H
End Function

be stuck in a room to play. Spatial freedom was a gift for the republic of players and so going mobile was their prior preference at that time.

Time Goes Forward, and so, the Consumer Preferences and Choices: By today's standards, just going mobile is not good enough, gamers want lighter laptops, more performance, high end graphic cards, high resolution/fps screens, mechanical keyboard, and the list is long ...

Today, gamers all over the planet become more demanding, their preferences changed drastically and so the industry does while trying to keep up with the human desires.

Back to our question, it is obvious that nobody will care about a 20 year old gaming laptop, even if it was a best seller with 1 million 5-star reviews at that time. Why is that? simply because it becomes **obsolete** by the **modern user criteria**. Its 1 million review doesn't matter anymore. And so, **as all the things and beings**, **reviews** also have **an expiration date**. where they become irrelevant to the buyer.

Although, a product, laptop, movie or hotel may had very good reviews once, but time took off their power, their importance and their effect over the judgement and decision of the consumer. At the end, **"you cannot beat time"**.

Algorithm 2 Review Time Score

Define: $RT_j = \{rt_{1j}, rt_{2j}, \ldots, rt_{nj}\}$: The set of reviews posting time expressed for the entity j.
Input: RT_j
Function T (rt_{ij}) **:**
  if $currentYear - rt_{ij} \geqslant 100$ **then**
    $T \longleftarrow 0.8$
  else
    $T \longleftarrow 1 - (currentYear - rt_{ij}) \times 0.002$
  end if
  return T
End Function

Algorithm 3 Review Sentiment Orientation Score

Define: $R_j = \{r_{1j}, r_{2j}, \ldots, r_{nj}\}$: The set of reviews expressed for the entity j.
$BERT_j = \{bert(r_{1j}), bert(r_{2j}), \ldots, bert(r_{nj})\}$: The set of output vectors of fine-tuned $BERT_{Base}$ (the sentiment orientation probability of reviews expressed for the entity j.
Input: R_j
Function S (r_{ij}) **:**
  $S \longleftarrow max(bert(r_{ij}))$
  return S
End Function

To conclude, we believe that more recent reviews generally provide users with more up-to-date information. Therefore, we design formula (2) to assign a time score to each review.

$$T(r_{ij}) = \begin{cases} 0.8 & \text{if } y - rt_{ij} \geqslant 100 \\ 1 - (y - rt_{ij}) \times 0.002 & \text{Otherwise} \end{cases} \quad (2)$$

We denote:
$T(r_{ij})$: Time score of review r_{ij}.
rt_{ij}: Publication year of review r_{ij}.
y: Current year.

The time score for a review ranges between 0.8 and 1, which implies that a higher time score is assigned to the most recent reviews.

Algorithm 2 computes the time score for review r_{ij}.

With the help of a film critic, we have been able to determine suitable minimum values for each of the review helpfulness and review time scores. Indeed, using different minimum values for both scores as parameters, multiple experiments have been performed on a various range of movies, where in each one among these, we compare the generated reputation value to the film critic's own rating regarding a given movie. Which leads to 0.75 and 0.8 to be chosen successively as the fittest minimum values for review helpfulness and review time scores. Next, given the high accuracy achieved through our reputation system, the same last experiments have been done on other domains such as products, restaurants and services, where we noticed very good results, particularly when using 0.75 and 0.8 as the minimum values for each of the scores.

E. REVIEW SENTIMENT ORIENTATION

We fine-tuned *BERT* model to determine the sentiment orientation probability of a target review due to the fact that it has achieved state-of-the-art results in a wide variety of natural language processing tasks by learning contextual relations between words or sub-words in a text. In this paper, we have interest in assigning a sentiment orientation score to each review. Since fine-tuned BERT returns an array of 2 values: probability of being negative and probability of being positive (Softmax activation function), we apply the *max* function to the fine-tuned BERT output vector. The highest probability is kept as the sentiment orientation score of the target review.

$$S(r_{ij}) = max([P_{r_{ij}}^{negative}, P_{r_{ij}}^{positive}]) \quad (3)$$

We denote:
$S(r_{ij})$: Sentiment orientation score for review r_{ij}.
$P_{r_{ij}}^{negative}$: BERT model output prediction for review r_{ij} being negative.
$P_{r_{ij}}^{positive}$: BERT model output prediction for review r_{ij} being positive.

The sentiment polarity of a target review r_{ij} is predicted as negative if $P_{r_{ij}}^{negative} > P_{r_{ij}}^{positive}$ and predicted as positive if $P_{r_{ij}}^{negative} < P_{r_{ij}}^{positive}$.

Algorithm 3 computes the sentiment orientation score for review r_{ij}.

F. REVIEW SCORE

Based on the above scores, we design formula (4) to compute a numerical score for each review:

$$RS(r_{ij}) = \frac{H(r_{ij}) + T(r_{ij}) + S(r_{ij})}{3} \quad (4)$$

We denote:
$RS(r_{ij})$: Review score for review r_{ij}.
$H(n_{ij})$: Helpfulness score of review r_{ij}.
$T(r_{ij})$: Time score of review r_{ij}.
$S(r_{ij})$: Sentiment orientation score for review r_{ij}.

Since review helpfulness score, review time score and review sentiment orientation score range between 0 and 1, the generated review score is also between 0 and 1.

Algorithm 4 computes the review score for all reviews.

Table 2 represents an example results of review score.

G. REPUTATION GENERATION

We propose formula (5) to compute a single reputation value toward the target entity using review score $RS(r_{ij})$ and review attached rating v_{ij}:

$$Rep(E_j) = \frac{\sum_{i=1}^{O_j} RS(r_{ij}).v_{ij}}{O_j} \quad (5)$$

Algorithm 4 Review Score

Define : $R_j = \{r_{1j}, r_{2j}, \ldots, r_{nj}\}$: The set of reviews expressed for the entity j.
$RH_j = \{rh_{1j}, rh_{2j}, \ldots, rh_{nj}\}$: The set of reviews helpfulness votes expressed for the entity j.
$RT_j = \{rt_{1j}, rt_{2j}, \ldots, rt_{nj}\}$: The set of reviews posting time expressed for the entity j.
$RS_j = \{rs_{1j}, rs_{2j}, \ldots, rs_{nj}\}$: The set of reviews score expressed for the entity j.
Input : R_j, RH_j, RT_j
Output: RS_j

for i *in range*(n) **do**
  $rs_{ij} \longleftarrow (H(rh_{ij}) + T(rt_{ij}) + S(r_{ij}))/3$
end for

TABLE 2. Example results of review score.

Review	Review helpfulness score	Review time score	Review sentiment orientation	Review sentiment score	Review score
Review 1	1	0.968	Positive	0.99732805	0.98844268
Review 2	0.75	0.982	Positive	0.99679191	0.9095973
Review 3	0.87468842	0.974	Negative	0.99608659	0.94825834
Review 4	0.91100877	0.964	Positive	0.9970323	0.95734702
Review 5	0.77448754	0.96	Negative	0.99694509	0.91047754

Algorithm 5 Reputation Generation

Define : $V_j = \{v_{1j}, v_{2j}, \ldots, v_{nj}\}$: The set of ratings expressed for the entity j.
$RS_j = \{rs_{1j}, rs_{2j}, \ldots, rs_{nj}\}$: The set of reviews score expressed for the entity j.
Rep: reputation value toward the target entity j.
Input : V_j, RS_j
Output: *Rep*

$temp \longleftarrow 0$
for i *in range*(n) **do**
  $temp \longleftarrow temp + rs_{ij} \times v_{ij}$
end for
$Rep \longleftarrow temp/n$

We denote:

E_j: Target entity j.

$Rep(E_j)$: Reputation value toward the target entity j.

$RS(r_{ij})$: Review score for review r_{ij}.

v_{ij}: Attached numerical rating to review r_{ij}.

O_j: Total number of reviews expressed for the target entity j.

The reputation value varies from 1 to 5 or 1 to 10 depending on the target entity attached rating values range.

Algorithm 5 computes the reputation value toward a target item.

Assuming that an entity E_j contains three reviews where $RH_j = \{100, 50, 1\}$, $RT_j = \{2020, 2010, 2000\}$, $BERT_j = \{0.998, 0.997, 0.996\}$ and $V_j = \{10, 10, 10\}$. By applying formula (1) and (2), we get the helpfulness and time scores: $H(r_{1j}) = 1$, $H(r_{2j}) = 0.849$, $H(r_{3j}) = 0.75$, $T(r_{1j}) = 1$, $T(r_{2j}) = 0.98$ and $T(r_{3j}) = 0.96$. After applying formula (4), we get the reviews scores: $RS_j = \{0.999, 0.942, 0.902\}$. In order to compute the reputation value toward E_j, we need to compute the product of rs_{ij} and v_{ij}. We get $rs_{1j}.v_{1j} = 9.99$, $rs_{2j}.v_{2j} = 9.42$ and $rs_{3j}.v_{3j} = 9.02$. Since $Rep(E_j) = \frac{\sum_{i=1}^{O_j} rs_{ij}.v_{ij}}{O_j}$, we can conclude that the first review r_{1j} has the highest impact (the highest product 9.99) on the reputation value of the entity E_j since it is very helpful and recent. In the contrary, the third review r_{3j} has the lowest impact (the lowest product 9.02). In fact, while it has the same attached rating as the first review, but, being both unhelpful and old made it by far less influential.

To conclude, recent and helpful reviews have more impact on the reputation value than old and unhelpful ones.

H. REPUTATION VISUALIZATION

It is important to provide a potential customer or user with sufficient information for the purpose of assisting his decision. Thus, we propose a new way to visualize reputation by depicting the produced numerical reputation value toward the target entity, opinion categories, positive review with the highest review score (formula 4) and negative review with the highest review score (Figure 6).

V. EXPERIMENT RESULTS

A. EXPERIMENTAL DATA COLLECTION AND PREPROCESSING

Five miscellaneous domains were addressed in our experiments, movie, TV show, product, hotel, and restaurant. We collected user reviews from IMDb,[13] TripAdvisor[14] and Amazon[15] using a web scraping tool called ScrapeStorm.[16] We extracted 400 reviews for 4 movies, 400 reviews for 4 TV shows, 200 reviews for 2 products, 100 reviews for 1 hotel and 100 reviews for 1 restaurant. Each extracted review contains: raw text, review time, review helpfulness votes and review attached rating (Figure 4). The statistical information of dataset is shown in Table 3.

TABLE 3. Statistical information of dataset.

Domain	Number of entities	Number of reviews	Number of reviews per entity
Movie	4	400	100
TV show	4	400	100
Product	2	200	100
Hotel	1	100	100
Restaurant	1	100	100

After collecting the reviews, we:

1) lowercase our text since we are using a BERT lowercase model
2) tokenize it
3) break words into WordPieces

[13]https://www.imdb.com/
[14]https://www.tripadvisor.com/
[15]https://www.amazon.com
[16]https://www.scrapestorm.com/

TABLE 4. BERT-base model classification result on Large Movie Review Dataset v1.0.

	Precision	Recall	F1 score	Accuracy
BERT-Base	0.88048	0.89816	0.88923204	0.88812

TABLE 5. Comparison result on Large Movie Review Dataset v1.0.

Approach	Accuracy
Vanilla CNN	80.35
Vanilla LSTM	80.72
Vanilla BiLSTM	81.73
BERT base	**88.81**

4) map our words to indexes using a vocab file that BERT provides
5) add special "CLS" and "SEP" tokens
6) append "index" and "segment" tokens to each input

B. SENTIMENT ANALYSIS

We fine-tune $BERT_{Base}$ model to predict the sentiment orientation of the collected reviews. We build the model by creating a single new layer that will be trained with Large Movie Review Dataset v1.0 [64][17] which contains 25,000 positive and 25,000 negative processed movie reviews. We set the sequence lenght to 128, the batch size to 32, the learning rate to 0.00002 and the number of epochs to 3. Table 4 depicts the performance of fine-tuned BERT-base model on Large Movie Review Dataset v1.0.

We compared BERT-Base model with Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (BiLSTM). GloVe embeddings were used to train LSTM, BiLSTM and CNN. Table 5 depicts the performance of fine-tuned BERT-base model, Vanilla CNN, Vanilla LSTM and Vanilla BiLSTM on Large Movie Review Dataset v1.0.

We can see from Table 5 that BERT-Base model achieves the highest sentiment analysis accuracy compared to Vanilla CNN, Vanilla LSTM and Vanilla BiLSTM.

We have mentioned in the literature review section some successful pre-trained models that achieve state-of-the-art results on Large Movie Review Dataset v1.0 such as XLNet-Large that achieves an accuracy of 96.21 and BERT-Large that achieves an accuracy of 95.79. However, these models require the combination of GPUs with plenty of computing power and a massive amount of memory.

We test BERT-Base model on our collected dataset. Figure 5 represents the accuracy of the model in predicting sentiment orientation for the collected reviews.

We observe from Figure 5 that the model achieves good results in predicting the sentiment polarity on the extracted reviews. Even more impressively, the model performs well on dataset 9, 10, 11 and 12 that contain product, hotel and restaurant reviews despite the fact that it's trained with movie reviews.

[17]https://ai.stanford.edu/~amaas/data/sentiment/

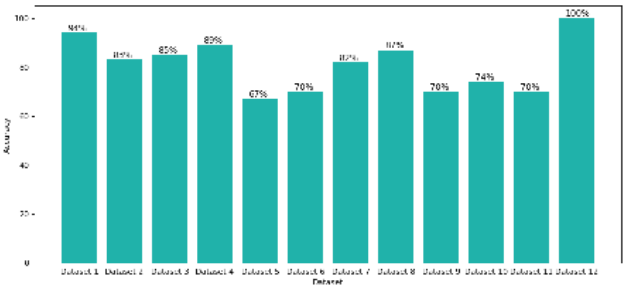

FIGURE 5. BERT-Base sentiment classification accuracy of the collected reviews.

TABLE 6. Comparison results: reputation visualization.

Work	Opinion categories	Top positive review	Top negative review
Yan et al. (2017) [5]	✓	✗	✗
Benlahbib & Nfaoui (2019) [9]	✗	✗	✗
Benlahbib & Nfaoui (2019) [10]	✓	✗	✗
This study	✓	✓	✓

C. REPUTATION VISUALIZATION

We propose a new way to visualize reputation by depicting the produced numerical reputation value toward the target entity, opinion categories, top positive review and top negative review.

As illustrated in Figure 6, our system provides users and potential customers with a reputation visualization form that shows the numerical reputation value toward the target entity, opinion categories (very good, good, neutral, very bad and bad) in a pie chart, top positive review (positive review that holds the highest review score) and top negative review (negative review that holds the highest review score).

Compared to previous studies on reputation generation [5], [9], [10], our proposed system is the only one that presents all of these helpful information in order to support users and customers during the decision making process in e-commerce websites. Table 6 shows comparison results between our system and previous reputation systems in term of reputation visualization.

D. SYSTEM EVALUATION

Previous studies on reputation generation based on mining user and customer reviews expressed in natural language have mainly focused on exploiting semantic and sentiment relations between reviews to generate a single reputation value toward various entities. However, customer and user reviews contain a lot of other useful information that could be exploited during the reputation generation phase like review posting time and review helpfulness votes. Unfortunately, up-to-date, no work has incorporated review time, review helpfulness votes and review sentiment polarity to produce a single numerical reputation value. Therefore, we propose a new reputation system that combines review posting time, review helpfulness votes and review sentiment orientation

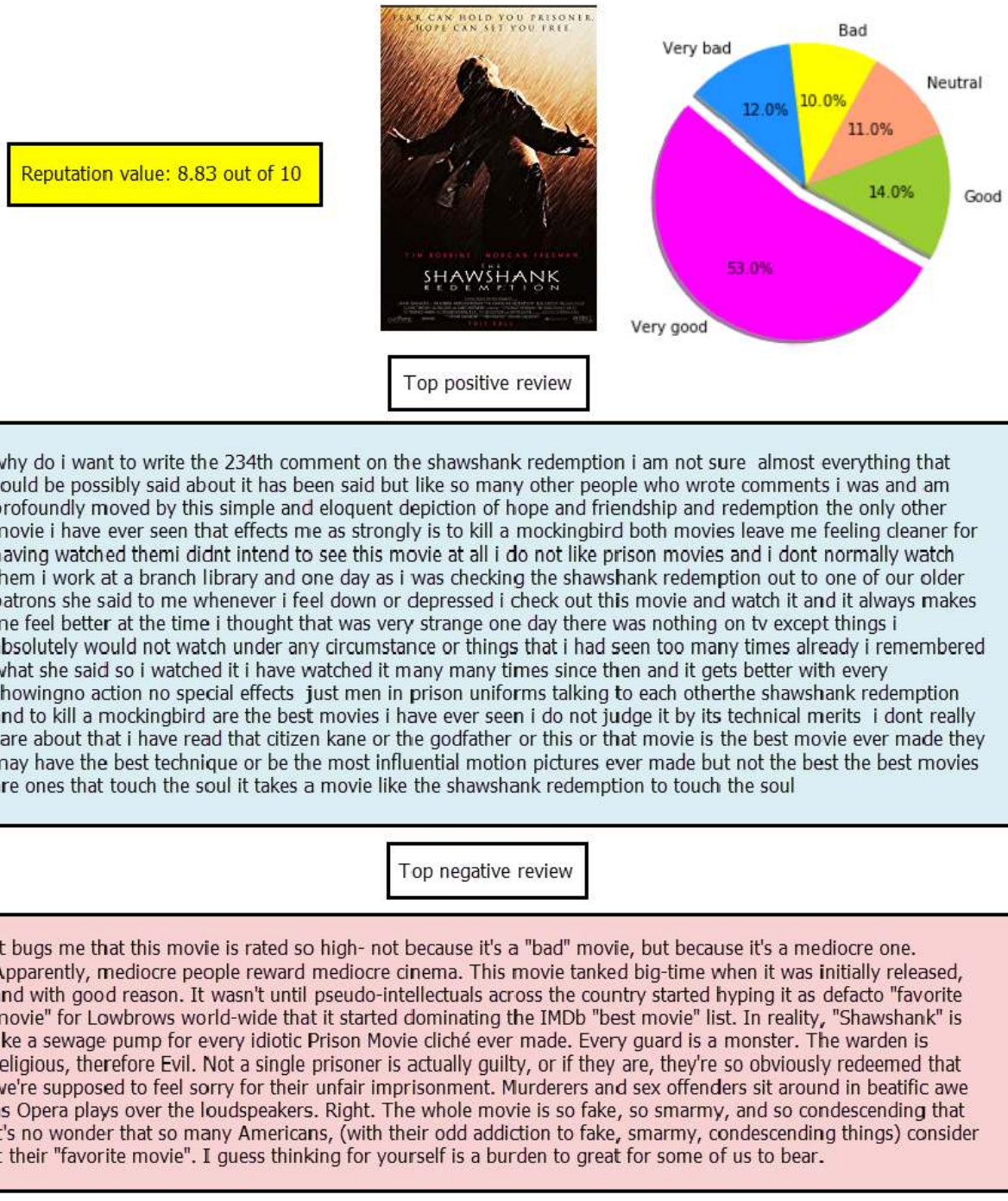

FIGURE 6. Reputation visualization.

in order to generate an accurate and reliable reputation value toward different entities. Table 7 depicts the difference between previous reputation systems [5], [9], [10] and our proposed reputation system.

Since there are no standard evaluation metrics to assess the effectiveness and robustness of reputation systems, we conduct a user and expert survey as adopted in many research papers [65]. We have invited 32 users and 3 experts to rate four reputation generation systems: System 1 (our reputation system), system 2 [5], system 3 [9] and system 4 [10]. Each user and expert assigns a satisfaction score to each reputation system. The score is ranged between 1 and 10.

The 32 users are from different backgrounds: 6 computer science PhD students, 2 math PhD students, an electrical engineer, an undergraduate student in mathematics, 2 computer science engineers, a physics teacher, 4 mathematics

TABLE 7. Comparison results: review attributes exploited by recent reputation systems.

Work	Semantic	Sentiment	Review helpfulness	Review time
Yan et al. (2017) [5]	✓	✗	✗	✗
Benlahbib & Nfaoui (2019) [9]	✓	✗	✗	✗
Benlahbib & Nfaoui (2019) [10]	✓	✓	✗	✗
This study	✓	✓	✓	✓

teachers, a research engineer in computer science, an electronic engineering student, an information systems engineer, a third year student at the National School of Commerce and Management, a quality control technician, a sixth year medical student, a housewife, 7 second year medical students and a software engineer.

Table 8 presents the average satisfaction scores for each reputation system given by the thirty users.

The formula of the average satisfaction score is: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ where $\{x_1, x_2, \ldots, x_N\}$ are the observed values of the sample items and N is the number of observations in the sample. The standard deviation is a measure of the amount of variation or dispersion of a set of values [66]. The formula for the standard deviation is: $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$ where $\{x_1, x_2, \ldots, x_N\}$ are the observed values of the sample items, μ is the mean value of these observations, and N is the number of observations in the sample.

We can see from Table 8 that 31 users favor our reputation system over the three other systems in term of helpfulness and effectiveness in generating reputation and visualization since it achieves the highest average satisfaction scores and the lowest standard deviation of satisfaction scores. Moreover, only one user (user 19) favors system 2 [9]. System 2 takes the second place by achieving an average satisfaction scores of 7.83. System 4 [5] comes next with a 7.01 average satisfaction scores, which sounds very reasonable since the main goal of system 2 was to improve system 4 by exploiting both sentiment and semantic analysis techniques. System 3 [9] takes the last place by achieving an average satisfaction scores of 5.625. System 3 doesn't provide users and customers with sufficient information to support their decision since providing reputation value alone isn't enough to help them make a judgment about a target item, the customers need more helpful information that could support them during their decision making process such as opinion categories, top positive review and top negative review.

We enrich our experiment results by inviting 3 experts to rate each reputation system with a satisfaction score. Expert 1 is a former owner of an e-commerce website whose main field of interest is natural language processing and machine learning, while expert 2 is an active e-commerce buyer and seller with more than 8 years of experience. As for expert 3, he is a second year PhD student in economics sciences.

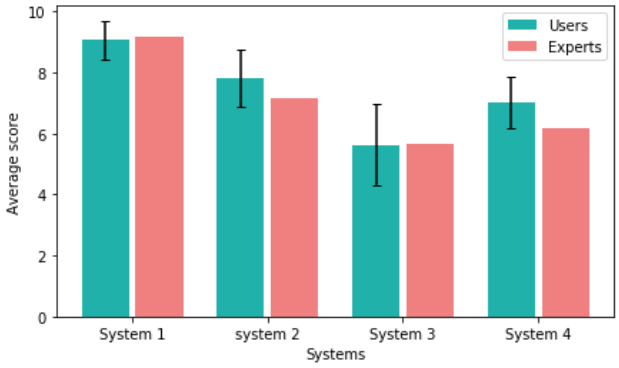

FIGURE 7. Users and experts average satisfaction scores for each reputation system.

Table 9 presents the average satisfaction scores for each reputation system given by the three experts.

Based on the average satisfaction scores given by the three experts (Table 9), reputation system 1 takes the first place with an average satisfaction scores of 9.17, preceded by system 2, system 4, and system 3 comes in last place with 5.67 as average satisfaction scores.

Figure 7 combines the results of Table 8 and Table 9.

Figure 7 shows that both users and experts choose system 1 as the best in term of reputation generation and visualization. system 2 holds the second place, preceded by system 4. system 3 comes at fourth place.

We asked the three experts to share their opinions about system 1 strengths and weaknesses. Table 10 contains expert reviews toward system 1.

E. FURTHER DISCUSSION

In summary, our reputation system exhibits the following advantages:

- **Accuracy**: The system incorporates review helpfulness, review time, review sentiment orientation probability and review attached rating in order to generate an accurate reputation value.
- **Holistic**: The system proposes a new form of reputation visualization that depicts numerical reputation value, opinion categories, top positive review and top negative review. The system also has the ability to output the top-k positive reviews and the top-k negative reviews. This new form of reputation visualization provides customers with sufficient information toward the target item in order to make a decision (buying, renting, booking) toward it.
- **Generality**: The system can be applied in any website that allows web users to: (1) post their reviews expressed in natural languages, (2) share their numerical or star ratings and (3) vote for helpful reviews. Furthermore, the system can be applied on various domains (products, movies, services, hotels).
- **Usefulness**: The system is very useful in term of supporting web customers during their decision making

TABLE 8. User satisfaction comparison.

Systems	Ours	Benlahbib & Nfaoui [10]	Benlahbib & Nfaoui [9]	Yan el al. [5]
User 1	**9**	8.5	6	8
User 2	**9.5**	8	7.5	6.9
User 3	**9**	8.5	8	8
User 4	**9.5**	8	4.5	8
User 5	**8**	6.5	6	6.5
User 6	**9.5**	8	5.5	7
User 7	**9**	7	4	7
User 8	**9**	8	7	8
User 9	**9**	7	5	7
User 10	**9**	6	4	6
User 11	**9**	8	5.5	8
User 12	**9**	8	6	7
User 13	**8**	6.5	5	6.5
User 14	**8**	5	3	5
User 15	**9.5**	8.5	5.5	7
User 16	**8**	7	4	7
User 17	**9.5**	8	7	7.5
User 18	**9**	8.5	6	7.5
User 19	8	**9**	2	5
User 20	**8**	7	6.5	7
User 21	**9**	**9**	6	6
User 22	**9**	8	6	7.5
User 23	**10**	8	5	6
User 24	**9**	7	5	6
User 25	**10**	8.5	6	7
User 26	**10**	9	8	8.5
User 27	**10**	8	5	7
User 28	**9**	**9**	7	8
User 29	**9**	8	6	7
User 30	**9.75**	8	5	6.5
User 31	**9**	8	6	7
User 32	**10**	9	7	8
Average	**9.07**	7.83	5.625	7.01
Standard Deviation	**0.63**	0.93	1.32	0.84

TABLE 9. Expert satisfaction comparison.

Systems	Ours	Benlahbib & Nfaoui [10]	Benlahbib & Nfaoui [9]	Yan el al. [5]
Expert 1	**10**	6	5	6
Expert 2	**8.5**	7.5	5	5.5
Expert 3	**9**	8	7	7
Average	**9.17**	7.17	5.67	6.17

process in E-commerce by instantly providing them with sufficient information toward the target item, saving them from spending both their time and effort on reading thousands of online reviews.

However, our reputation system suffers from:

- **Safety**: Due to the openness of Internet, many malicious users post fake reviews (false positive/false negative) aiming to impact the popularity and credibility of online

TABLE 10. Expert reviews.

Expert	Review
Expert 1	The present work proposes a new method for reputation generation and visualization by incorporating helpfulness votes and time review features to sentiment and semantic features. In addition, the system outputs the item reputation value, top positive review, top negative review and a pie chart which shows the distribution of sentiment over the reviews. Based on all these properties, i give a 10 to system 1. I think that this work can be enhanced by mining the top reviews according to the system's user. For instance, if two reviews have been selected as top positive review, what review should the system display? By analyzing the users data and behaviour, the system will output the most accurate review with regard to user preferences.
Expert 2	I choose system 1 as the best because it covers many important criteria neglected by other systems. On the one hand, the numerical value and opinion categories alone only reveal the general impact of the film (product) on users (good film, nice film, bad film) and ignore the in-depth details that are essential for any kind of reputation. On the other hand, System 1 takes into account all four attributes of the evaluation, which will have a positive effect on the accuracy and credibility of the reputation value. I give it 8.5/10, because in my opinion, the method still has some shortcomings since the best positive/negative reviews can contain spoilers for movies or bad personal experiences for a service, and this would create some confusion for the user.
Expert 3	I underline the importance of this study in terms of its usefulness and interest. It is an issue that seeks to put in place a system that allows subjective analysis, and thus scrutinizes and makes the opinions of viewers more concrete. And so, after having an idea of the different reputation systems, I consider the first system to be the most efficient in terms of reputation generation since it takes into consideration more variables/determinants such as review helpfulness votes and review time.

products. Therefore, our system should incorporate a filtering phase in order to detect and remove fake and irrelevant reviews.

VI. CONCLUSION

In this paper, we have proposed a reputation system that generates reputation toward various items (products, movies, TV shows, hotels, restaurants, services) by mining customer and user reviews expressed in natural language. The system incorporates four review attributes: review helpfulness, review time, review sentiment polarity and review rating. The system also provides a holistic reputation visualization form by depicting the numerical reputation value, opinion group categories, top positive review and top negative negative. To better evaluate the effectiveness of our reputation system, 32 users and 3 experts were invited to assign a score of one (least satisfaction) to ten (highest satisfaction) to four reputation generation systems. Our reputation system achieved the highest average satisfaction scores given by both users and experts. The three experts were also invited to share their point of view toward the proposed system in term of reputation generation and visualization.

We believe that the proposed system represents an interesting online reputation system, full of fascinating insights into customer's decision-making process in e-commerce web sites.

Future studies will focus on:

- exploiting further features such as user credibility (prolific reviewers) and user's online behavior as suggested by expert 1 (Table 10).
- detecting and removing fake and irrelevant reviews by applying a filtering phase and therefore reducing the processing time and increasing the efficiency of the system at once since only relevant and useful reviews will be taken into account.
- incorporating aspect based opinion mining during the phase of reputation generation and visualization. As a result, the reputation visualization will be enhanced. Indeed, the system will depict more useful information toward the target entity E such as its features ($E_{featureX}$, $E_{featureY}$, $E_{featureZ}$ ...), the number of positive reviews toward feature $E_{featureX}$, and the number of negative reviews toward feature $E_{featureY}$ ...

ACKNOWLEDGEMENT

A sincere thank you to Mohammed El Moutaouakkil (mohammed.elmoutaouakkil@usmba.ac.ma) for his diligent proofreading of this paper.

REFERENCES

[1] T. Hou, B. Yannou, Y. Leroy, and E. Poirson, "Mining customer product reviews for product development: A summarization process," *Expert Syst. Appl.*, vol. 132, pp. 141–150, Oct. 2019.

[2] Y.-H. Hu, Y.-L. Chen, and H.-L. Chou, "Opinion mining from online hotel reviews—A text summarization approach," *Inf. Process. Manage.*, vol. 53, no. 2, pp. 436–449, Mar. 2017, doi: 10.1016/j.ipm.2016.12.002.

[3] K. Bafna and D. Toshniwal, "Feature based summarization of customers' reviews of online products," *Procedia Comput. Sci.*, vol. 22, pp. 142–151, Jan. 2013.

[4] L. Zhuang, F. Jing, and X.-Y. Zhu, "Movie review mining and summarization," in *Proc. 15th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, New York, NY, USA: ACM, 2006, pp. 43–50, doi: 10.1145/1183614.1183625.

[5] Z. Yan, X. Jing, and W. Pedrycz, "Fusing and mining opinions for reputation generation," *Inf. Fusion*, vol. 36, pp. 172–184, Jul. 2017, doi: 10.1016/j.inffus.2016.11.011.

[6] A. Kangale, S. K. Kumar, M. A. Naeem, M. Williams, and M. K. Tiwari, "Mining consumer reviews to generate ratings of different product attributes while producing feature-based review-summary," *Int. J. Syst. Sci.*, vol. 47, no. 13, pp. 3272–3286, Oct. 2016, doi: 10.1080/00207721.2015.1116640.

[7] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA: ACM, 2004, pp. 168–177, doi: 10.1145/1014052.1014073.

[8] S. Pecar, "Towards opinion summarization of customer reviews," in *Proc. ACL, Student Res. Workshop*, Melbourne, VIC, Australia: Association for Computational Linguistics, Jul. 2018, pp. 1–8. [Online]. Available: https://www.aclweb.org/anthology/P18-3001

[9] A. Benlahbib and E.-H. Nfaoui, "An unsupervised approach for reputation generation," in *Proc. 2nd Int. Conf. Intell. Comput. Data Sci.*, vol. 148. Amsterdam, The Netherlands: Elsevier, 2019, pp. 80–86.

[10] A. Benlahbib and E. H. Nfaoui, "A hybrid approach for generating reputation based on opinions fusion and sentiment analysis," *J. Organizational Comput. Electron. Commerce*, pp. 1–19, Aug. 2019, doi: 10.1080/10919392.2019.1654350.

[11] A. Benlahbib, A. Boumhidi, and E. H. Nfaoui, "A logistic regression approach for generating movies reputation based on mining user reviews," in *Proc. Int. Conf. Intell. Syst. Adv. Comput. Sci. (ISACS)*, Dec. 2019, pp. 1–7.

[12] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th Int. Conf. Very Large Data Bases (VLDB)*, San Francisco, CA, USA: Morgan Kaufmann, 1994, pp. 487–499. [Online]. Available: http://dl.acm.org/citation.cfm?id=645920.672836

[13] G. A. Miller, "WordNet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995, doi: 10.1145/219717.219748.

[14] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *Proc. Conf. Human Lang. Technol. Empirical Methods Natural Lang. Process. (HLT)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 339–346, doi: 10.3115/1220575.1220618.

[15] L. Garcia-Moya, H. Anaya-Sanchez, and R. Berlanga-Llavori, "Retrieving product features and opinions from customer reviews," *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 19–27, May 2013.

[16] A. Abbasi, S. France, Z. Zhang, and H. Chen, "Selecting attributes for sentiment classification using feature relation networks," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 3, pp. 447–462, Mar. 2011.

[17] X. Chi, T. P. Siew, and E. Cambria, "Adaptive two-stage feature selection for sentiment classification," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2017, pp. 1238–1243.

[18] B.-G. Sasha, H. Kerry, M. Ryan, N. Tyler, R. George, and R. Jeff, "Building a sentiment summarizer for local service reviews," in *Proc. WWW Workshop NLP Inf. Explosion Era*, 2008, pp. 339–348.

[19] A. M. El-Halees and D. Salah, "Feature-based opinion summarization for arabic reviews," in *Proc. Int. Arab Conf. Inf. Technol. (ACIT)*, Nov. 2018, pp. 1–5.

[20] O. Ahlgren, "Research on sentiment analysis: The first decade," in *Proc. IEEE 16th Int. Conf. Data Mining Workshops (ICDMW)*, Dec. 2016, pp. 890–899.

[21] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Inf. Fusion*, vol. 36, pp. 10–25, Jul. 2017, doi: 10.1016/j.inffus.2016.10.004.

[22] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 79–86, doi: 10.3115/1118693.1118704.

[23] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[24] M. E. Maron, "Automatic indexing: An experimental inquiry," *J. ACM*, vol. 8, no. 3, pp. 404–417, Jul. 1961, doi: 10.1145/321075.321084.

[25] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguistics*, vol. 22, no. 1, pp. 39–71, 1996. [Online]. Available: http://dl.acm.org/citation.cfm?id=234285.234289

[26] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," *Comput. Intell.*, vol. 22, no. 2, pp. 110–125, May 2006, doi: 10.1111/j.1467-8640.2006.00277.x.

[27] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proc. 42nd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Barcelona, Spain, Jul. 2004, pp. 271–278. [Online]. Available: https://www.aclweb.org/anthology/P04-1035

[28] M. Koppel and J. Schler, "The importance of neutral examples for learning sentiment," *Comput. Intell.*, vol. 22, pp. 100–116, May 2006. [Online]. Available: https://www.microsoft.com/en-us/research/publication/the-importance-of-neutral-examples-for-learning-sentiment/

[29] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann, 1993.

[30] X. Yan and X. G. Su, *Linear Regression Analysis*. Singapore: World Scientific, 2009. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/6986

[31] E. Frank and M. Hall, "A simple approach to ordinal classification," in *Proc. 12th Eur. Conf. Mach. Learn. (EMCL)*. London, U.K.: Springer-Verlag, 2001, pp. 145–156. [Online]. Available: http://dl.acm.org/citation.cfm?id=645328.649997

[32] S. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proc. ACL, 50th Annu. Meeting Assoc. Comput. Linguistics*. Jeju Island, South Korea: Association for Computational Linguistics, Jul. 2012, pp. 90–94. [Online]. Available: https://www.aclweb.org/anthology/P12-2018

[33] R. Jing, Y. Yu, and Z. Lin, "How service-related factors affect the survival of B2T providers: A sentiment analysis approach," *J. Organizational Comput. Electron. Commerce*, vol. 25, no. 3, pp. 316–336, Jul. 2015.

[34] Ł. Augustyniak, P. Szyma ski, T. Kajdanowicz, and W. Tuligłowicz, "Comprehensive study on lexicon-based ensemble classification sentiment analysis," *Entropy*, vol. 18, no. 1, p. 4, Dec. 2015.

[35] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[36] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006.

[37] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. Int. Conf. Mach. Learn.*, vol. 96, 1996, pp. 148–156.

[38] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *J.-Jpn. Soc. Artif. Intell.*, vol. 14, nos. 771–780, p. 1612, 1999.

[39] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: Understanding rating dimensions with review text," in *Proc. 7th ACM Conf. Recommender Syst. (RecSys)*, New York, NY, USA: ACM, 2013, pp. 165–172, doi: 10.1145/2507157.2507163.

[40] P. D. Turney, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 417–424, doi: 10.3115/1073083.1073153.

[41] P. D. Turney, "Mining the Web for synonyms: PMI-IR versus LSA on TOEFL," in *Proc. 12th Eur. Conf. Mach. Learn. (EMCL)*, London, U.K.: Springer-Verlag, 2001, pp. 491–502. [Online]. Available: http://dl.acm.org/citation.cfm?id=645328.650004

[42] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Comput. Linguistics*, vol. 37, no. 2, pp. 267–307, Jun. 2011. [Online]. Available: https://www.aclweb.org/anthology/J11-2001

[43] S. Vashishtha and S. Susan, "Fuzzy rule based unsupervised sentiment analysis from social media posts," *Expert Syst. Appl.*, vol. 138, Dec. 2019, Art. no. 112834.

[44] M. Fernández-Gavilanes, T. Álvarez-López, J. Juncal-Martínez, E. Costa-Montenegro, and F. Javier González-Castaño, "Unsupervised method for sentiment analysis in online texts," *Expert Syst. Appl.*, vol. 58, pp. 57–75, Oct. 2016.

[45] R. Moraes, J. F. Valiati, and W. P. G. Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN," *Expert Syst. Appl.*, vol. 40, no. 2, pp. 621–633, Feb. 2013, doi: 10.1016/j.eswa.2012.07.059.

[46] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. 31st Int. Conf. Int. Conf. Mach. Learn. (ICML)*, vol. 32, 2014, pp. II–1188–II–1196. [Online]. Available: http://dl.acm.org/citation.cfm?id=3044805.3045025

[47] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Red Hook, NY, USA: Curran Associates, 2013, pp. 3111–3119. [Online]. Available: http://dl.acm.org/citation.cfm?id=2999792.2999959

[48] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Denver, CO, USA: Association for Computational Linguistics, May/Jun. 2015, pp. 103–112. [Online]. Available: https://www.aclweb.org/anthology/N15-1011

[49] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.*, vol. 36, no. 4, pp. 193–202, Apr. 1980.

[50] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[51] C. M. Bishop, *Neural Networks for Pattern Recognition*. New York, NY, USA: Oxford Univ. Press, 1995.

[52] G. Mesnil, T. Mikolov, M. Ranzato, and Y. Bengio, "Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews," 2014, *arXiv:1412.5335*. [Online]. Available: https://arxiv.org/abs/1412.5335

[53] K. Baktha and B. K. Tripathy, "Investigation of recurrent neural networks in the field of sentiment analysis," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Apr. 2017, pp. 2047–2050.
[54] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
[55] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," 2019, *arXiv:1904.12848*. [Online]. Available: http://arxiv.org/abs/1904.12848
[56] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805
[57] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*. [Online]. Available: http://arxiv.org/abs/1907.11692
[58] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," 2019, *arXiv:1906.08237*. [Online]. Available: http://arxiv.org/abs/1906.08237
[59] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," 2019, *arXiv:1909.11942*. [Online]. Available: http://arxiv.org/abs/1909.11942
[60] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified Text-to-Text transformer," 2019, *arXiv:1910.10683*. [Online]. Available: http://arxiv.org/abs/1910.10683
[61] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–14. [Online]. Available: https://openreview.net/forum?id=r1xMH1BtvB
[62] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, San Francisco, CA, USA, Tech. Rep., 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
[63] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, New Orleans, LA, USA: Association for Computational Linguistics, vol. 1, Jun. 2018, pp. 2227–2237. [Online]. Available: https://www.aclweb.org/anthology/N18-1202
[64] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proc. 49th Annu. Meet. Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Portland, OR, USA: Association for Computational Linguistics, Jun. 2011, pp. 142–150. [Online]. Available: http://www.aclweb.org/anthology/P11-1015
[65] D. Wang, S. Zhu, and T. Li, "SumView: A Web-based engine for summarizing product reviews and customer opinions," *Expert Syst. Appl.*, vol. 40, no. 1, pp. 27–33, Jan. 2013, doi: 10.1016/j.eswa.2012.05.070.
[66] J. M. Bland and D. G. Altman, "Statistics notes: Measurement error," *BMJ*, vol. 312, no. 7047, p. 1654, 1996, doi: 10.1136/bmj.313.7059.744.

ABDESSAMAD BENLAHBIB received the master's degree in computer science. He is currently pursuing the Ph.D. degree with the Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco. His research interests concern the applications of natural language processing (NLP) techniques for decision making in E-commerce websites.

EL HABIB NFAOUI (Member, IEEE) received the Ph.D. degree in computer science from Sidi Mohamed Ben Abdellah University, Fez, Morocco, and the University of Lyon, France, through a Cotutelle agreement (doctorate in joint-supervision), in 2008, and the HU Diploma degree (accreditation to supervise research) in computer science from Sidi Mohamed Ben Abdellah University, in 2013. He is currently a Professor of computer science with Sidi Mohamed Ben Abdellah University. He has published in international reputed journals, books, and conferences, and has edited seven conference proceedings and special issue books. His current research interests include information retrieval, language representation learning, machine learning and deep learning, Web mining and text mining, semantic Web, Web services, social networks, and multi-agent systems. He is a Co-Founder and an Executive Member of the International Neural Network Society Morocco Regional Chapter. He co-founded the International Conference on Intelligent Systems and Computer Vision (ISCV), in 2015, and the International Conference on Intelligent Computing in Data Sciences (ICSD), in 2017. He is the Co-Founder and the Chair of the IEEE Morocco Section Computational Intelligence Society Chapter. He has served as a Reviewer for scientific journals and on program committees of several conferences.