Received: June 29, 2021. Revised: August 25, 2021.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Leveraging Lexicon-Based and Sentiment Analysis Techniques for Online Reputation Generation

Achraf Boumhidi[1]* **El Habib Nfaoui[1]**

[1]*Computer Science Department, LISAC Laboratory, Faculty of Sciences Dhar EL Mahraz, Sidi Mohamed Ben Abdellah University, Morocco*
Corresponding author's Email: achraf.boumhidi@usmba.ac.ma

Abstract: Advanced reputation generation systems analyze user-generated content such as opinions and reviews expressed in natural language in order to produce a reliable and trusted reputation value. According to recent and relevant literature, while many important attributes are exploited such as semantic and sentiment orientation of the reviews, time, and opinions' relevancy, other relevant attributes such as users' credibility, as well as the sentiment intensity of the reviews, are not considered. In this paper, we propose a system that computes a single numerical reputation value between 0 and 10 from Twitter microblogging platform by incorporating the sentiment orientation of the tweets, the sentiment intensity of the positive tweets, as well as the users' and tweets' credibility score. To assess the effectiveness of our system we have compared its computed reputation value with the ground truth one ranging from 0 to 10. This later is the weighted average votes of thousands of users taken from IMDb, Amazon, TripAdvisor, and Yelp concerning respectively four products and services (movie: 7.9, phone product: 7.3, hotel: 9, restaurant: 7.1). The experimental results conducted on four real-world Twitter datasets related to the aforementioned products show that our system provides a reputation value (movie: 7.66, phone product: 7.18, hotel: 8.71, restaurant: 7.08) that is near to the ground truth one. Consequently, it can be applied in practice and used by consumers and businesses to generate a reliable and trusted reputation value from tweets in order to support them during their decision-making process in E-commerce platforms (buying, renting, booking, etc.).

Keywords: Reputation generation, Decision support system, Social media analysis, E-business applications, Natural language processing, Sentiment analysis.

1. Introduction

In the last decade, social media has taken the world by storm, and its popularity has grown dramatically. A survey of 1000 consumers revealed that approximately 55 percent of customers choose to call out brands on social media, and 4 in 5 consumers think social media has increased accountability for businesses[1]. With platforms like Twitter, millions of consumers now have a voice to share their opinions and experiences about products and services in form of short text. This large amount of people's opinions is very convenient for an automatic computing of a reputation value that defines customer satisfaction towards a specific item. Previous studies have mostly focused on generating reputation from e-commerce websites based on semantic and sentiment orientation of textual reviews, as well as time and opinions relevancy. However, social media platform particularly Twitter offers other useful information about users' credibility and relevancy of the opinions that could be extracted and analyzed. Every opinion shared by the user in Twitter consists of features that describe the instance of the user and the instance of the tweet[2]. Some features are "number of followers, which implies that the opinion shared by the user could be reached and consumed by his followers ", "account authenticity, it means if the user has a verified account, it indicates that the user is trusted and it lends credibility to his opinions", "number of

[1] https://sproutsocial.com/insights/data/q3-2017/

[2] https://developer.twitter.com/en/docs

likes (favorites), which implies that tweets that receive higher likes from other users, generally provide more information", and "number of retweets, which implies that the more retweets, the higher number of people will see the user's opinion". A reliable reputation generation system tends to incorporate all of those features of both instances, in order to produce an honest reputation that represents the quality of a business output. Furthermore, an advanced reputation generation system should extract more than just the sentiment orientation from the textual reviews. Analyzing the sentiment intensity of those reviews by differentiating between reviews with the same sentiment orientation based on the type of words expressed by the user would help produce a credible reputation value. In this paper, we proposed a reputation generation system able to compute a reputation value through the noisiness of the informal language expressed by users on Twitter. First, we applied a Bidirectional Encoder Representations from Transformers (BERT) classifier to extract the sentiment orientation of the textual opinions. Next, we calculated a sentiment intensity score by classifying positive reviews into three categories (low positive, medium positive, high positive) using semantic and syntactic techniques. Finally, we incorporate the previous results with a computed credibility score from the extracted Twitter features (number of followers, account authenticity, number of likes, number of retweets) to generate a single numerical reputation value (between 0 and 10) toward various entities (products, services, movies, hotels, restaurants).

The rest of this paper is organized as follows. Section 2 presents the related work. Section 3 presents the preliminaries. Section 4 describes our proposal. Section 5 details the experiments. Section 6 presents the discussion. Finally, Section 7 concludes this paper.

2. Related work

In this section, we present the literature review of reputation generation systems, as well as the recent document-level sentiment analysis techniques considering that our approach generates the reputation based on sentiment analysis.

2.1 Reputation generation systems

Reputation is the subjective qualitative belief a person has regarding a brand, person, company, product, or service[3]. It can be based on feelings, past experiences, and the viewpoint of a circle of "trusted" people. Reputation generation systems tend to compute a reputation value from user-generated content expressed online. There are two types of reputation generation systems. The first type is described as a simple reputation system, where it is based on statistical data such as the numerical ratings given to a specific entity by the users. In [1], they describe a simple technique to compute reputation scores, which is to sum the number of positive ratings and negative ratings separately and to keep a total score as the positive score minus the negative score. This is the principle used in eBay's reputation forum. Amazon used a slightly more advanced scheme proposed in [2], which calculate the reputation score as the average of all ratings. Advanced models in this category compute a weighted average of all the ratings, where the rating weight can be determined by factors such as the age of the rating, distance between rating and current score, etc. The second type of reputation generation system is referred to as an advanced reputation system, it is based on opinions and reviews expressed in natural language. Authors in [3] proposed the first reputation generation system based on textual opinions. Those opinions are filtered to eliminate unrelated ones and then grouped into a number of fused principal opinion sets that contain opinions with a similar or the same attitude or preference. Latent Semantic Analysis (LSA) model and cosine similarity have been used to compute the similarity between opinions before grouping them into different sets. By aggregating the ratings attached to the fused opinions, they normalize the reputation of an entity. In [4], authors have improved the previous work introduced in [3]. They proposed a hybrid approach that separates reviews into positive and negative based on their sentiment polarity by applying the two classifiers Naïve Bayes and Linear Support Vector Machine (LSVM). Then they grouped positive and negative reviews into principal opinion sets based on their semantic relations. Next, they calculate a custom reputation value separately for positive and negative groups by considering some statistics of principal opinion sets. Finally, they compute the final reputation value using weighted arithmetic mean. Another reputation generation system was proposed in [5] where they separate movie reviews collected from E-commerce websites into two groups: positive and negative based on their sentiment orientation using the Logistic Regression classifier. Then, they computed a custom reputation value by considering some statistics of each one of the groups. Finally, they used the weighted arithmetic

[3] https://blog.reputationx.com/whats-reputation

mean to compute the reputation value towards the target movie. In [6], authors proposed a new reputation generation system where they converted the ratings given by the users to textual words based on the intensity of the ratings, then they fused the output with the reviews. Next, they extract the sentiment polarity of the fused movie reviews using bidirectional long short-term memory (Bi-LSTM) classifier. Finally, they computed the reputation score based on the classification results. In [7], authors proposed an advanced reputation system that generates reputation toward various entities (products, movies, TV shows, hotels, restaurants, services) by mining customer reviews expressed in e-commerce websites. The system incorporates four review attributes: review helpfulness, review time, review sentiment polarity, and review rating. First, they designed two equations to compute review helpfulness and review time scores, and they fine-tuned a Bidirectional Encoder Representations from Transformers (BERT) model to predict the review sentiment orientation probability. Second, they designed a formula to assign a numerical score to each review. Finally, they proposed a new formula to compute reputation value toward the target entity. Experimental results using several real-world datasets of different domains collected from IMDb, TripAdvisor, and Amazon websites confirm the effectiveness of the proposed method in generating and visualizing reputation compared to the state-of-the-art reputation systems. Authors in [8] proposed a new reputation generation system that incorporates fine-grained opinion mining and semantic analysis to generate reputation toward movies and TV shows. Differently from previous studies on reputation generation that treat the task of document-level sentiment analysis as a binary classification problem (positive, negative), the proposed system identifies the sentiment strength during the phase of sentiment classification by using fine-grained sentiment analysis where reviews are classified into five classes: strongly negative, weakly negative, neutral, weakly positive and strongly positive. They first started by separating reviews into groups based on their sentiment orientation. Next, a custom score is computed for each opinion group. Finally, a numerical reputation value is produced toward the target movie or TV show. Experimental studies showed that the proposed system outperforms the reputation system proposed in [3] since it produces the nearest reputation values to the ground truth (IMDb weighted average ratings) for both movies and TV shows.

The various challenges of opinion mining and sentiment analysis have motivated us to propose an effective and reliable reputation generation system, which is tailored to handle the production of a trustworthy reputation value from opinions expressed on Twitter.

2.2 Document level sentiment analysis

Sentiment analysis (SA) is the computational study of people's opinions, appraisals, attitudes, and emotions toward entities [9]. It is the interpretation and classification of emotions within text data using text analysis techniques. There are three primary characterization levels in SA: document-level, sentence-level, and aspect-level. Considering that our approach is based on document-level sentiment analysis to extract the sentiment orientation of people's reviews on Twitter. This section will take aim at previous research work done on the area of document-level SA.

The Sentiment Analysis at the document level analyzes the text from a given document and indicates its overall sentiment polarity: positive, negative, or neutral. Various approaches have been used to tackle the task of document-level sentiment analysis:

Machine learning approaches: There is a wide variety of machine learning tasks and successful applications. The machine learning approach relies on the famous machine learning algorithms to solve the sentiment extraction as a normal text classification problem that employs syntactic or linguistic features[4]. In the paper [10], authors build a sentiment classifier based on the multinomial Naive Bayes classifier [11] that uses N-gram and POS-tags as features. This classifier can determine the sentiment orientation (positive, negative, or neutral) of the tweets collected using Twitter API[5]. They tested the impact of an n-gram order on the classifier's performance. The best performance is achieved when using bigrams. They explain it as bigrams provide a good balance between coverage (unigrams) and an ability to capture the sentiment expression patterns (trigram). In [12], authors implemented and compared the results of two Naive Bayes unigram models, a Naive Bayes bigram model, and a Maximum Entropy model [13] for tweets classification. They found that Naive Bayes classifiers worked much better than Maximum Entropy model. Authors in [14] suggested a solution

[4] https://medium.com/@safdar.mirza94/sentiment-analysis-machine-learning-approach-2adb57a1af91

[5] https://developer.twitter.com/en/docs/twitter-api

for sentiment analysis on Twitter data by applying distant supervision, in which the training data consisted of tweets with emoticons which served as noisy labels. They build models using Naive Bayes, Maximum Entropy, and Support Vector Machines (SVM) [15]. The feature space consisted of unigrams, bigrams, and POS. They concluded that SVM outperformed other models and that unigram was more effective as features. However, in [16], authors presented variations of Naive Bayes classifiers to extract the polarity of English tweets. Two different variants of Naive Bayes classifiers were built namely Baseline (trained to classify tweets as positive, negative, and neutral), and Binary (makes use of a polarity lexicon and classify into positive and negative. neutral tweets neglected). The features considered by classifiers were lemmas (nouns, verbs, adjectives, and adverbs), polarity lexicons, and multiword from different sources, and valence shifters. The results show that there is a performance improvement when the classifiers are implemented with the binary strategy, when they use a polarity lexicon, and when multi-words are considered as features. In [17], authors used an ensemble framework for sentiment classification acquired by employing various feature sets and classification techniques. In their work, they used two types of feature sets (Part-of-speech information and word relations) and three base classifiers (Naive Bayes, Maximum Entropy, and Support Vector Machines). They applied ensemble approaches like (fixed combination, weighted combination, and meta-classifier combination) for sentiment classification and obtained better accuracy.

Deep Learning approaches*:* Deep learning has emerged as a powerful machine learning technique that learns multiple layers of representations or features of the data and produces state-of-the-art prediction results in sentiment analysis. In [18] they made a practical comparison between Support Vector Machines (SVM) and Artificial Neural Networks (ANN) [19] for document-level sentiment classification, which demonstrated that ANN produced competitive results to SVM's in most cases. To overcome the weakness of bag-of-words, researchers in [20] proposed Paragraph Vector, an unsupervised learning algorithm that learns vector representations for variable-length texts such as sentences, paragraphs, and documents. The vector representations are learned by predicting the surrounding words in contexts sampled from the paragraph. They achieved new state-of-the-art results on several sentiment analysis tasks. In this paper [21], they employed a convolutional neural network (CNN) [22] on text categorization to exploit the 1D structure of text data for accurate prediction. Instead of using low-dimensional word vectors (Word2vec [23], Glove [24]) as input, they directly apply CNN to high-dimensional text data, which leads to directly learning embedding of small text regions for use in classification. The experiments demonstrate the effectiveness of the approach in comparison with state-of-the-art methods on the IMDb dataset[6]. In this paper [25], they propose a novel approach to detect emotions like happy, sad, or angry in textual conversations using an LSTM [26] based deep learning model. The approach consists of semi-automated techniques to gather training data for the model. they exploit the advantages of semantic and sentiment-based embeddings (Glove) and propose a solution combining both. Evaluation of the model on real-world textual conversation outperforms CNN and LSTM baselines as well as other machine learning baselines. In [27], Google researchers developed a Bidirectional Encoder Representation from Transformers which is a state-of-the-art machine learning model used for NLP tasks including sentiment analysis. Unlike previous language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. BERT obtained new state-of-the-art results on eleven natural language processing tasks such as (GLUE[7], MultiNLI[8], SQuAD[9]). Authors in [28] introduced FinBERT, a language model based on BERT, to tackle natural language processing (NLP) tasks in the financial sentiment analysis domain. They achieved state-of-the-art results by a significant margin on two evaluation datasets. In addition to BERT, they also implemented other pre-training language models like ELMo [29] and ULMFit [30] for comparison purposes. ULMFit, further pre-trained on a financial corpus, beat the previous state-of-the-art for the classification task, only to a smaller degree than BERT. In [31] Facebook AI and University of Washington researchers figured ways to improve Google's BERT language model by introducing a robustly optimized BERT approach (RoBERTa). RoBERTa relies on pretraining with larger batches of data and changes to the masking pattern of training data. While in pretraining, the original BERT uses masked language modeling and next-sentence prediction, but RoBERTa drops the next-sentence prediction

[6] https://www.imdb.com/interfaces/
[7] https://gluebenchmark.com
[8] https://cims.nyu.edu/~sbowman/multinli
[9] https://rajpurkar.github.io/SQuAD-explorer

approach. RoBERTa exceeded state-of-the-art results in GLUE benchmark dataset. Recently researchers at Carnegie Mellon University and Google AI Brain Team proposed XLNet [32], a generalized autoregressive pretraining method that enables learning bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order and overcomes the limitations of BERT thanks to its autoregressive formulation. When trained on a very large NLP corpus, the model achieves state-of-the-art performance for the standard NLP tasks including sentiment analysis. In [33] google researchers proposed Text-to-Text Transfer Transformer (T5). The model was pre-trained on C4[10] and it achieved state-of-the-art results on many NLP benchmarks while being flexible enough to be fine-tuned to a variety of important downstream tasks such as Sentiment classification. The T5 model achieved an accuracy of 97.4% on the SST-2 binary classification dataset.

3. Preliminaries

This section focuses on outlining the problem statement, introducing BERT architecture used in our reputation generation system, and describing the Twitter network and its features.

3.1 Problem statement

This paper deals with the problem of generating a single numerical reputation value ∈ [0,10] for products or services from tweets by incorporating a score generated from sentiment polarity of the tweets, sentiment intensity score in positive tweets, as well as a users' and tweets' credibility score generated based on the extracted relevancy features: number of followers, account authenticity, number of retweets, number of likes. Relying on a set of textual reviews (tweets) $T_j = \{t_{1j}, t_{2j}, \dots, t_{nj}\}$ expressed for an entity E_j, the set of sentiment orientation of the tweets T_j predicted by $BERT_{base}$ classifier $BERT_j = \{\text{Bert}(t_{1j}), \text{Bert}(t_{2j}), \dots, \text{Bert}(t_{nj})$ where $\text{Bert}(t_{ij}) \in [0,1]$, the set of sentiment intensity score in positive tweets $SE_j = \{se_{1j}, se_{2j}, \dots, se_{nj}\}$, and finally, the relevancy features extracted for each element in the T_j set described in Table 1. The goal is to compute a sentiment polarity score SPS_j based on the calculated percentage of positive tweets from the sentiment orientation set $BERT_j$, then calculate the sentiment intensity score SI_j from the SE_j set, and the credibility score CS_j from the elements of Table 1 in order to generate a final reputation value Rep_j by aggregating all the previous outputs.

Table 1. User and tweet relevancy features

Feature	Description	Notation
Number of followers	Number of people following the user.	N_f
Account authenticity	If the author of the tweet has a verified account.	A_{ver}
Number of retweets	Number of retweets received for a particular tweet.	N_{rt}
Number of likes	Number of likes (favorites) received for a particular tweet.	N_L

3.2 Bidirectional encoder representations from transformers

Bidirectional Encoder Representations from Transformers (BERT) is a Transformer-based [34] machine learning technique for natural language processing (NLP) developed by Google. It is a bidirectional trained model since it makes use of a technique called Masked Language Modeling (MLM): it randomly masks words in the sentence and then it tries to predict them. Masking means that the model looks in both directions and it uses the full context of the sentence, both left, and right surroundings, to predict the masked word. Context-free models like word2vec generate a single word embedding representation (a vector of numbers) for each word in the vocabulary. However, context-based models like BERT generate a representation of each word that is based on the other words in the sentence. BERT is also trained on a next sentence prediction task to better handle tasks that require reasoning about the relationship between two sentences. There are two main types of pre-trained versions of BERT depending on the scale of the model architecture, Table 2 shows the differences between the two BERT models.

BERT outperformed the state-of-the-art across a

Table 2. Comparison between BERT-Base and BERT-Large

	Layers	Hidden Nodes	Attention Heads	Parameters
BERT-Base	12	786	12	110M
BERT-Large	24	1024	16	340M

[10] https://www.tensorflow.org/datasets/catalog/c4

Table 3. Twitter relationships between users and tweets

	Tweet	User
User	Like	Follow
	Retweet	Retweet to
	Reply to	Reply to
	Post	Mention

wide variety of tasks under general language understanding including sentiment analysis. In the remainder of the paper, we will be using the BERT-Base pre-trained model to predict the sentiment orientation of our textual tweets.

3.3 Twitter network description

Twitter is a well-known online social network and a microblogging service on which users post and interact with messages of up to 280 characters known as "tweets". These tweets or messages are public by default and visible to all those who are following the tweeter. The Twitter network is not bidirectional, meaning that the connections don't have to be mutual: you can follow users who don't follow you back and the other way around[11]. Each tweet can also have replies from other people creating real-time conversations around hot topics, new products, and interesting new content. In Twitter, when a user 'A' is following another user 'B', we say that 'A' is a follower of 'B'. Three main types of actions can be made between two users. The user can replay to another user through one of his tweets. Also, a user can share, or retweet the tweet of another user. And finally, a user can like or mark a tweet of another user as a favorite. Table 3 describes the relationship between users and tweets.

4. Proposed approach

4.1 System overview

Our approach is divided into several steps. After collecting a dataset of tweets reflecting the opinions of the users about a specific entity, the first step of our reputation generation system is to clean and process the tweets. The following step allows calculating a sentiment polarity score based on the results of the sentiment classification of those tweets. Next, a sentiment intensity score is calculated based on the classification results of only the positive tweets using semantic and syntactic techniques. Finally, a credibility score is calculated using a formula that

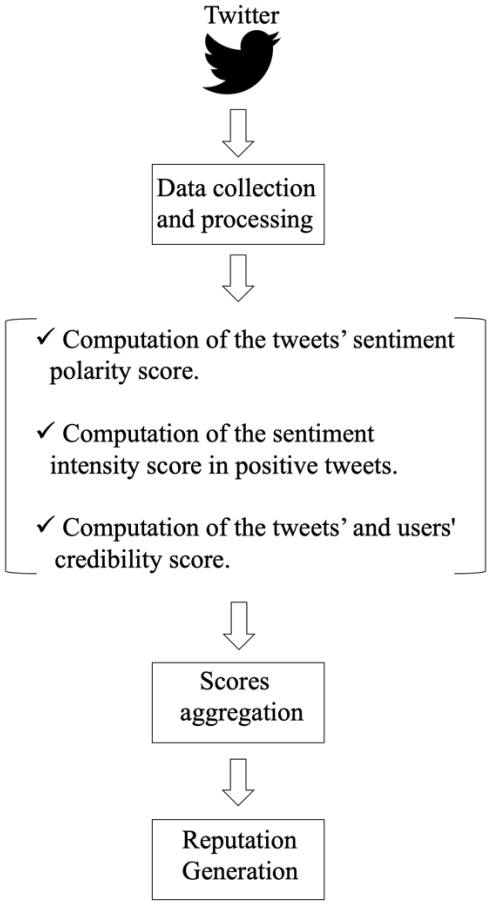

Figure. 1 Proposed system pipeline

incorporates the extracted users' and tweets' features: number of followers, account authenticity, number of retweets, and number of likes. Our system aggregates the computed scores and generates a reputation value about the specific entity. Fig. 1 describes the pipeline of our reputation generation system.

4.2 Data collection and processing

With the use of Twitter API[12] and python scripts, we were able to retrieve data from the blogging site Twitter, considering it as a source of information in our approach. Our reputation generation system exploits more than just the textual content of the tweets to generate a reputation value. In addition, other statistical features are scraped from the Twitter platform such as the number of followers, account authenticity, number of retweets, and number of likes. Fig. 2 shows the features exploited in our approach extracted from an opinion expressed on Twitter. The collection of the textual reviews is processed and cleaned, also non-standard text spelling words are

[11] https://www.socialmediatoday.com/content/twitter-101-what-twitter-really-about

[12] https://developer.twitter.com/en/docs/twitter-api

Figure. 1 Features considered for the proposed approach.

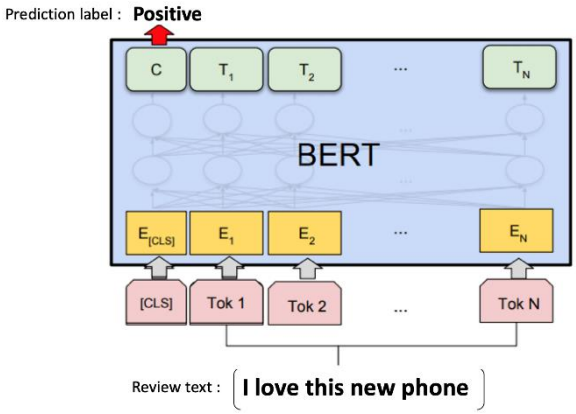

Figure. 2 BERT-Base binary sentiment classifier architecture

handled in order to eliminate noises associated with the raw informal text. The pre-processing steps applied in this paper are as follows:

- Replacing slang words.
- Special characters and punctuation marks are removed.
- Eliminating tweets containing advertising and publicity links.

Technical and statistical details of the data collection and processing are described in section 5.

4.3 Sentiment polarity score

Algorithm 1: Review sentiment orientation prediction

Define:

$T_j = \{t_{1j}, t_{2j}, \ldots, t_{nj}\}$: the set of textual reviews expressed for a specific entity j.

$BERT_j = \{\text{Bert}(t_{1j}), \text{Bert}(t_{2j}), \ldots, \text{Bert}(t_{nj})\}$: the set of the output of the BERT-Base sentiment classifier using the Sigmoid function for the entity j.

Input: T_j
Function $SO\ (t_{ij})$:
  if $Bert(t_{ij}) > 0.5$ then
    $SO \leftarrow 1$
  else
    $SO \leftarrow 0$
  end if
  return SO
End Function

After processing and cleaning the reviews, we fine-tuned the BERT-based model and applied it to perform the sentiment polarity prediction of the given reviews. Every review in the dataset is provided to the BERT classifier in order to classify it into two possible categories (positive and negative). Fig 3 shows an example of a classification model architecture used in this paper.

The input representation used by BERT can represent our sentence in a single sequence of tokens. The first token of every input sequence is the special classification token [CLS], this token is used in classification tasks, and it is taken as the fixed-dimensional pooled representation of the input sequence. The label probabilities are computed with a standard Sigmoid. Using the sigmoid as the activation function, the predictions formula will give us different values of probabilities between zero and one. Following machine learning conventions, every time we have a probability value bigger than "0.5" we assign the value "1" (positive) to the prediction label. On the other hand, every time we have a probability value lower than "0.5" we assign the value "0" (negative) to the prediction label, Algorithm 1 describes this step.

The sentiment polarity score 'SPS_j' is calculated using formula Eqs. (1) and (2) based on the set of sentiment orientation $SO_j = \{so_{1j}, so_{2j}, \ldots, so_{nj}\}$ expressed for the entity 'j' predicted by algorithm 1.

$$SPS_j = \frac{P \times 10}{T} \quad (1)$$

Where:

$$T = P + N \quad (2)$$

We denote:
P: Total number of positive reviews in the SO_j set.
N: Total number of negative reviews in the SO_j set.
T: Total number of positive and negative reviews.

4.4 Sentiment intensity score

As we explained earlier in the paper, every textual review is classified into two possible categories (positive and negative). However, in this section, our goal is to classify every review with a positive sentiment orientation into three classes (low positive, medium positive, high positive) by applying the entity-level sentiment analysis using the lexicon-based approach. Since we are dealing with reviews and opinions about products and services, there is a difference between just liking a product, and wanting to buy the product. We believe that a review where a user expresses his purchase intention and the willingness to pay for the product/service is more important than just liking it, and would impact positively the reputation generated for that entity. On the other hand, the negative reviews will not go through the same process as the positive reviews, since the intensity of the adverse sentiment in the negative reviews is irrelevant as far as we are concerned. We have collected all the positive reviews obtained from the previous classification phase to create a dataset called the positive reviews dataset. Then we annotated every textual review from the dataset using basic Part-Of-Speech tagging (POS tags)[13], which is also called grammatical tagging**,** and it is the process of marking up a word in a text as corresponding to a particular part of speech. Parts of speech (POS) are specific lexical categories to which words are assigned, based on their syntactic context and role. Usually, words can fall into one of the following major categories (Verb, Adjective, Adverb)[14]. We manually created three sets. The first set contains some specific e-commerce related verbs that could be used by the user while expressing his opinion on Twitter, the second set of words contains some of the positive descriptive adjectives, and finally, the third set contains some of the adverbs of degree. Next, we developed a script that detects if the (verbs, adjectives, adverbs) stated in the textual review exist in the three sets, and a numerical value

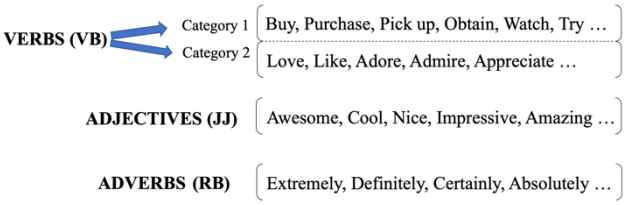

Figure. 3 An example of words in each set with their assigned score

is assigned to each of the words in the review based on which set the word belongs. Figure 4 shows some of the words in each set with their numerical value. We separated the words in the verbs set into two categories. The first one 'Cat1' contains verbs that confirm the buying intent of the user, which will have a higher score compared to the second category 'Cat2' contains verbs showing the admiration of the user for a specific product.

Every verb (VB), adjective (JJ), and adverb (RB) that compose the positive textual reviews expressed by the user will get a score based on the three previous sets described in Fig. 4. All nouns, pronouns, and prepositions in the textual review will be ignored and will have a score of zero. In case of a recurrence of verbs, adjectives, and adverbs in the review, we will only assign a score for the first one, the other duplicated words will have a score of zero. Now to classify the positive review into (low positive, medium positive, high positive), we sum those scores assigned to words and apply the following rules:

- If the sum of all the scores in the review is less than 3, then it's a low positive (LP).
- If the sum of all the scores in the review is equal to 3, then it's a medium positive (MP).
- If the sum of all the scores in the review is greater than 3, then it's a high positive (HP).

Fig. 5 shows the classification of three reviews using lexicon-based approach in order to determine the sentiment intensity of the positive reviews through e-commerce related terms and utterances used in Twitter.

The next step is to calculate a numerical value from this previous result that will influence our final reputation value. Therefore, after classifying all the positive reviews in our dataset we calculated the percentage of (LP, MP, HP). Next, we employed Formula Eq. (3) which will output a sentiment intensity score for the entire dataset that ranges

[13] https://en.wikipedia.org/wiki/Part-of-speech_tagging

[14] https://www.kdnuggets.com/understanding-language-syntax-and-structure-practitioners-guide-nlp-3.html

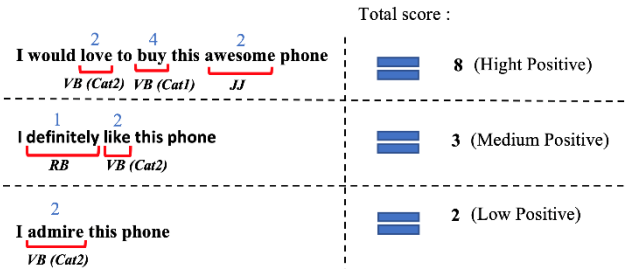

Figure. 4 An example of the classification of three positive reviews based on the sum of the assigned values to each word

Table 4. Chosen features with their assigned coefficients.

Feature	Weight
Number of followers ($\boldsymbol{N_f}$)	0.175
Account authenticity ($\boldsymbol{A_{ver}}$)	0.05
Number of Retweets ($\boldsymbol{N_{rt}}$)	0.175
Number of likes ($\boldsymbol{N_L}$)	0.1

between 0 and 0.5 and it will be used to compute the final reputation value for a specific entity. A higher SI_j score means the more positive the final reputation will be.

$$SI_j = \frac{Perc_{HP} - Perc_{LP}}{2} \quad (3)$$

We denote:
$Perc_{HP}$ *:* percentage of the high-positive reviews in our positive reviews' dataset.
$Perc_{LP}$ *:* percentage of the low-positive reviews in our positive reviews' dataset.

4.5 Tweet's and user's credibility score

We can find a lot of features that concern the opinions expressed on Twitter. In this work, we have chosen the best possible features that will help produce the most accurate reputation value. The features showed in Table 1 (number of followers, account authenticity, number of retweets, number of likes) are extracted with every given opinion. Our goal in this part of the paper is to calculate a numerical score using the combination of these four relevancy features. We started by identifying the max value of every statistical feature ($maxN_f$, $maxN_{rt}$, $maxN_L$) in the entire dataset except the account authenticity feature since it has a binary representation, we affect 1 if the user's account is verified and 0 if not. Moreover, we believe that some features have more weight than others. For example, the number of retweets is more important than the number of likes because by retweeting the opinion whether it's positive or negative it could reach and influence other users which will impact the reputation value. A Twitter study[15] of the top 150 destination marketers in the U.S. uncovered some interesting things about likes (favorites) and retweets. Both likes and retweets reflect the quality of engagement with followers, but the retweets can act as a platform for further distributing content and can translate to greater reach. By retweeting posts, users essentially endorse the content to all of their followers and thus tend to gravitate towards entertaining information over explicitly promotional materials. "Users must be excited to spread a message to their own followers in order to retweet a post". Also, according to a study conducted by researchers from Microsoft and Carnegie Mellon University, they learned that the most credible tweets come from people we follow [35]. Therefore, in order to distinguish between the importance of the features, we conducted experiments to determine the values of the weights that provide the best results when calculating the reputation value. We noticed very good results, particularly when using the weight shown in Table 4.

The next step is to calculate a numerical value for every positive and negative review in our dataset that indicates the relevancy of that review. Based on the relevancy features (N_f, A_{ver}, N_{rt}, N_L) and their assigned weights described in Table 4, we calculate a value R_j by applying the formula (4), the output is a numerical value between 0 and 0.5 that will be used to compute the credibility score CS_j.

$$R_j = \frac{N_f \times 0.175}{maxN_f} + \frac{N_{rt} \times 0.175}{maxN_{rt}} + \frac{N_L \times 0.1}{maxN_L} + A_{ver} \times 0.05 \quad (4)$$

We denote:
$\boldsymbol{N_f}$: Number of people following the user who expressed his review on Twitter.
$\boldsymbol{A_{ver}}$: If the author of the tweet has a verified account then A_{ver} equal to 1 if not A_{ver} equal to 0.
$\boldsymbol{N_{rt}}$: Number of retweets (re-posting) of user's review on Twitter.
$\boldsymbol{N_L}$: Number of likes received for a review expressed by the user.
$\boldsymbol{maxN_f}$: the max number of Followers of a user in the entire dataset.

[15] https://medium.com/favorites-vs-retweets-and-why-one-is-more-important-than-the-other

Table 5. An example of the relevancy features of four reviews.

	Number of Followers (N_f)	**Account Authenticity (A_{ver})**	**Number Of retweets (N_{rt})**	**Number Of Likes (N_L)**	**Sentiment Orientation**
Review 1	11000	1	3000	1000	Positive
Review 2	600	0	500	200	Negative
Review 3	1200	0	600	300	Negative
Review 4	25000	1	500	8500	Positive

$\boldsymbol{maxN_{rt}}$: the max number of retweets of a user's review in the entire dataset.

$\boldsymbol{maxN_L}$: the max number of likes received for a review in the entire dataset.

After calculating the value R_j for every review, the second step is to compute a score for the entire dataset called the credibility score CS_j that will be used to calculate the final reputation value. First, we grouped all the values R_j of the positive reviews in the set $R_j^{Positive}$ = { $R_{1j}^{Positive}$, $R_{2j}^{Positive}$, …, $R_{nj}^{Positive}$ }, and we grouped all the values R_j of the negative reviews in the set $R_j^{Negative}$ = { $R_{1j}^{Negative}$, $R_{2j}^{Negative}$, …, $R_{nj}^{Negative}$ }. Next, we employed the formula Eq. (5) in order to calculate the credibility score which is a numerical value between 0 and 0.5 that will be used to compute the final reputation value.

$$CS_j = \frac{\sum_{i=0}^{n} R_{ij}^{Positive} - \sum_{i=0}^{n} R_{ij}^{negative}}{T} \qquad n, T \in \mathrm{N} \tag{5}$$

With

$$T = \begin{cases} len\,(R_j^{positive}) & \text{if } \left(\sum_{i=0}^{n} R_{ij}^{Positive} - \sum_{i=0}^{n} R_{ij}^{negative}\right) \geq 0 \\ len\,(R_j^{negative}) & \text{otherwise} \end{cases}$$

We denote:

$R_j^{Positive}$: set of the values calculated in formula Eq. (4) of all the positive reviews.

$R_j^{Negative}$: set of the values calculated in formula Eq. (4) of all the negative reviews.

$len\,(R_j^{positive})$: The length of the positive set $R_j^{positive}$.

$len\,(R_j^{Negative})$: The length of the negative set $R_j^{positive}$.

Table 6. An example of the output of the formula Eq. (4) based on the relevancy features in Table 5

	'R' value
Review 1	0.313
Review 2	0.0352
Review 3	0.0469
Review 4	0.341

For a better understanding, we created a small dataset by collecting the relevancy features (N_f, A_{ver}, N_{rt} , N_L) of four reviews as shown in Table 5. The first step in our procedure is to identify the max values for every feature ($maxN_f$, $maxN_{rt}$, $maxN_L$) as explained earlier in the paper. Next, we apply the previous formula Eq. (4) in order to calculate the value 'R' for each review, the results are displayed in Table 6.

Now it's time to compute the credibility score CS_j of our small dataset based on the 'R' values showed in table 6 obtained from the formula Eq. (4). We separated those values into two sets based on their sentiment orientation: the positive set $R^{Positive}$, and the negative set $R^{Negative}$ Next, we applied the formula Eq. (5) to calculate the credibility score CS = 0.28595 that will be used to compute the final reputation value.

4.6 Reputation generation

Based on the preceding results, we are now able to generate a final reputation value of the entity 'j' using the sentiment polarity score SPS_j , the sentiment intensity score SI_j and finally, the credibility score CS_j by employing the following formula Eq. (6):

$$Rep_j = SPS_j + \mathrm{SI}_j + CS_j \tag{6}$$

- ***SPS***: a value than range between 0 and 10 and it is generated based on the results of BERT sentiment classifier using the formula Eq. (1).
- ***SI***: a value that range between 0 and 0.5 computed based on the classification of positive

Table 7. Statistical information about the evaluation datasets

	Domain	Name of the Entity	Number of Tweets
Dataset 1	Movie	The Irish man (2019)	1000
Dataset 2	Product	iPhone 12	1000
Dataset 3	Hotel	Caesars Las Vegas	1000
Dataset 4	Restaurant	In-N-out	1000

reviews, this value will only impact our final reputation positively since the sentiment intensity score can't be negative. Thus, if the positive reviews are intensely positive the final reputation value will increase with a max value of 0.5.

- ***CS:*** a value that range between -0.5 and 0.5 which is generated based on the users' and tweets' relevancy features, this score can affect the final reputation positively or negatively, knowing that an entity that has a higher credibility score would cause the reputation to increase and vice versa.

5. Experiment results

5.1 Experimental data collection and pre-processing

To assess our reputation generation system, we have collected four real-word datasets using Twitter API. Each dataset contains 1000 Tweets in which the users express their opinions and feedbacks about that specific entity, the datasets also contains the statistical data concerning the relevancy of the users and the tweets. For a higher evaluation efficiency, we have decided to use a different domain for each dataset (movie, product, hotel, restaurant). Table 7 displays statistical information about the datasets. Each extracted opinion contains: raw textual tweet, the number of retweets, the number of likes received for the tweet, the number of user's followers, and the user's account authenticity. Table 8 shows a sample of the evaluation dataset that contains reviews about the movie "The Irishman (2019) ".
Now after collecting the reviews, we started by:

- Eliminating links associated with the textual tweet.
- Lowercasing every word in the textual tweet.
- Removing special characters and punctuations.
- Preparing the cleaned textual tweet for the BERT classifier (tokenization, special tokens addition ...).

5.2 Sentiment classification evaluation

We fine-tuned the BERT-base-uncased model on the SST-2 binary classification dataset (consist of 9613 for the training set and 1821 for the test set)[16] using Pytorch Hugging Face's transformers library[17]. We took the pre-trained BERT model, add an untrained layer of neurons at the end, and train the new model for our sentiment classification. As for the parameters of the model, we used a learning rate of 2e-5 (0.00002), the batch size is 32, the sequence length equal to 64, and we used three epochs for our training phase. Table 9 displays the performance achieved of our BERT model on the SST2 dataset.

We compared our BERT-based model with the Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Bidirectional Long Short-Term Memory (BiLSTM) in order to evaluate our model's performance. We used pre-trained Glove embedding with 840B tokens, 2.2M vocab, 300d vectors in CNN, LSTM, and BiLSTM to map words into numerical vector spaces. Table 10 present a comparison of performance between the finetuned BERT model used in our approach and the other deep learning models on the SST-2 dataset

As shown in Table 10, the BERT-based model, despite being a simple architecture, performs better in terms of accuracy than the other models. However, we mentioned in literature the existence of other powerful deep learning models such as ELECTRA, BERT-Large, T5 with state-of-the-art results on the SST-2 dataset. The reason for not choosing these models is because they are computationally expensive with millions of parameters, which require the combination of GPUs with plenty of computing power and a massive amount of memory.

We performed the sentiment prediction using our BERT-based model on the four datasets collected. Fig. 6 shows the accuracy of the sentiment orientation of the reviews in each dataset. As we can see, all the datasets exceeded 90% in accuracy except the second dataset, but still, the model was able to achieve very good performance despite the fact that it was trained on a dataset containing reviews about movies (SST-2). Fig. 7 shows the results of the sentiment classifier on the evaluation datasets.

5.3 System evaluation

[16] https://nlp.stanford.edu/sentiment/

[17] https://huggingface.co/transformers/

Table 8. A sample of a user's opinion from Dataset 1

Raw Tweet	Number of Retweets	Number of Likes	Number of Followers	Account Authenticity
So, I just watched The Irish man for the first time. This movie is …my goodness! Off to watch it again	421	1911	32 027	1

Table 9. BERT-base performance on the SST-2 binary classification dataset

	F1 Score (%)	Accuracy (%)
BERT base	90.94	91.22

Table 10. Accuracy (%) of the proposed model on SST-2 dataset compared to other previous models

Mode	Accuracy (%) on the 'SST-2 Dataset'
RNN	82.2
LSTM	84.9
CNN	87.2
Bi-LSTM	87.5
BERT base	**91.2**

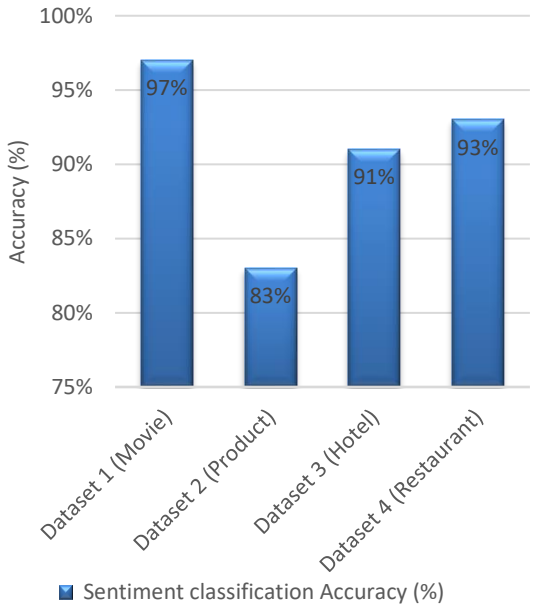

Figure. 5 Accuracy of the BERT-base sentiment classifier for the evaluation datasets

In order to evaluate the credibility of our system's output and the effectiveness of its components, and due to the nonexistence of standard evaluation metrics for this kind of system. We conducted a ground truth comparison and features comparison.

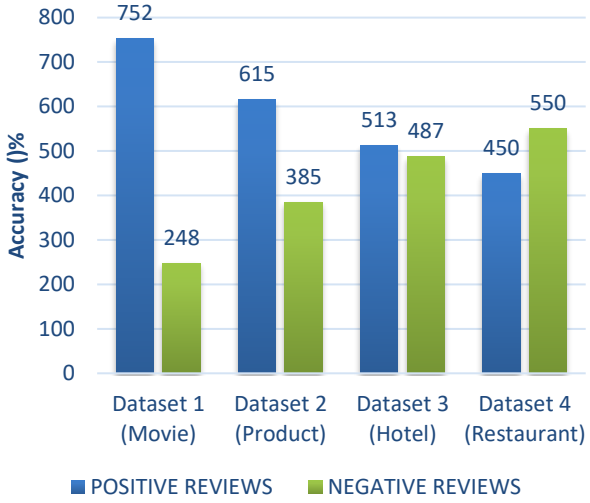

Figure. 6 Classification results of the positive and negative reviews in each evaluation dataset

5.3.1. Ground truth comparison

To evaluate the reputation values generated by our system for the evaluation datasets described in Table 7, we compared the computed reputation value with the weighted average votes provided by four crowdsourcing websites for each domain in our evaluation datasets. Those websites provide a platform to millions of users and experts to rate various entities by giving a numerical value between 0 and 10 that represents the users' satisfaction toward a specific item. The average of all those ratings is considered as a ground truth, that will be compared with our system's output. Figure 8 shows the IMDb users weighted average vote for The Irishman (2019) movie with the number of users who rated the movie. As we can see, more than three hundred thousand people rated the movie which makes this weighted average vote reliable as a ground truth for further comparison with our system's computed reputation value. Table 11 shows the websites (IMDb [18], Amazon[19], TripAdvisor[20], Yelp[21]) used as a ground truth for each domain in our evaluation datasets.

Our proposed system takes each evaluation

[18] https://www.imdb.com/
[19] https://www.amazon.com/
[20] https://www.tripadvisor.com/
[21] https://www.yelp.com/

Figure. 7 IMDb users weighted average vote for the movie "The Irishman (2019)"

Table 11. Weighted average vote for 4 entities in different domain

	Domain	Website Name (ground truth)	Number of ratings	Weighted average vote
Dataset 1: The Irish man (2019)	Movie	IMDb	318 157	7.9
Dataset 2: iPhone 12	Product	Amazon	17 314	7.3
Dataset 3: Caesars Las Vegas	Hotel	TripAdvi sor	26 246	9.0
Dataset 4: In-N-out	Restaur ant	Yelp	9413	7.1

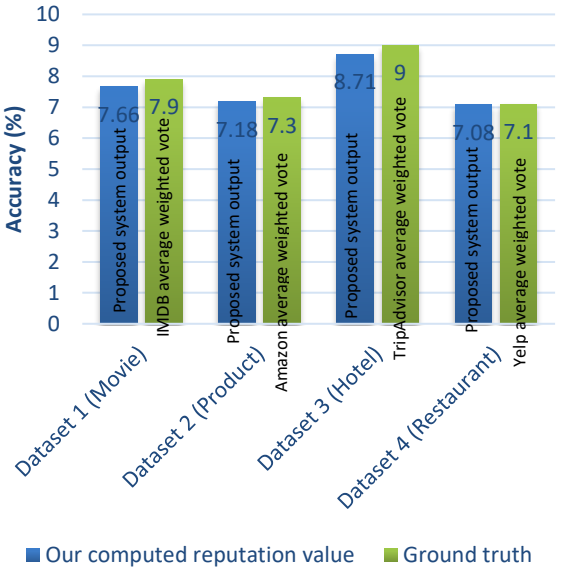

Figure. 8 Reputation value generated by the proposed system vs. ground truth average weighted vote

dataset as an input in order to compute a numerical reputation value that represents the user's feedback toward the target entity. The results are then compared to the weighted average vote given by IMDb, Amazon, TripAdvisor, and Yelp as demonstrated in Fig. 9. As we can see, our system was able to produce a reputation value from opinions expressed on Twitter that is very close to the ground truth values, and can be used in real-life applications

Table 12. Comparison results considering the number of users' ratings and tweets

	Weighted average vote (Ground Truth)	Our system's computed reputation value
Dataset 1	7.9 (318 157 users' ratings on IMDb)	**7.66** (1000 tweets)
Dataset 2	7.3 (17 314 users' ratings on Amazon)	**7.18** (1000 tweets)
Dataset 3	9 (26 246 users' ratings on TripAdvisor)	**8.71** (1000 tweets)
Dataset 4	7.1 (9413 users' ratings on Yelp)	**7.08** (1000 tweets)

to generate a reliable reputation value about various entities.

According to Table 12, we notice that our reputation generation system is able to produce a numerical value that is close to the ground truth for four different domains (movie, product, hotel, restaurant). However, reputation values produced by our proposed system were generated from only 1000 textual tweets compared with the weighted average vote (ground truth) that relies on a large number of users' ratings.

5.3.2. Features Comparison

Previous studies of reputation generation systems have essentially focused on generating a reputation value from e-commerce websites relying on semantic, sentiment analysis of the textual reviews, time, and opinion relevancy. However, in this paper, we proposed the first reputation generation system that generates a reputation value of a specific entity based on data extracted from the micro-blogging platforms such as Twitter, which contains and shares more challenging short text. Moreover, up-to-date no work took into consideration the credibility of the users expressing the opinions, and also the amount of people influenced by the opinions shared by the users. Also, no previous reputation generation system considered the sentiment intensity toward a specific entity shared by the users by distinguishing between a review where the user expresses his thoughts about simply liking the product and the willingness of consuming that product. Therefore, our reputation generation system incorporates the sentiment orientation of the reviews in addition to the users' and tweets' relevancy features and finally, the sentiment intensity in positive reviews to generate an accurate reputation value toward various types of entities. Table 13 outlines the differences between previous reputation generation systems and our proposed one. Our system employs new significant attributes that

Table 13. Features comparison between our proposed reputation generation system and previous state-of-the-art systems

System	Semantic	Sentiment	Time attribute	Opinion relevancy	Users credibility	Sentiment intensity
System 1: [3]	✓	✗	✗	✗	✗	✗
System 2: [4]	✓	✓	✗	✗	✗	✗
System 3: [7]	✓	✓	✓	✓	✗	✗
Our System	✓	✓	✓	✓	✓	✓

boost the credibility and reliability of the generated reputation value. As we can see in Table 13, our system included the credibility and relevancy of the users sharing the opinions based on features extracted from Twitter. Our system also takes into consideration the sentiment intensity of the user toward a specific entity. By incorporating all of the features presented in Table 13, our system was able to produce a reliable and trustworthy reputation value. Table 13. Features comparison between our proposed reputation generation system and previous state-of-the-art systems.

6. Discussion

The purpose of this work has been to generate a reputation value from opinions and reviews expressed on the social media platform "Twitter". Given that the reputation generation systems used in commercial and social media platforms have serious vulnerabilities, for that reason, the reliability of these systems sometimes is questionable. We proposed two basic criteria for judging the reliability of reputation computation systems.

(i) The sentiment intensity of the opinion: since we are generating a reputation score based on textual reviews expressed on Twitter, the pieces of information and the sentiment orientation extracted from these opinions heavily affect that reputation score. Most of the reputation generation systems only consider the binary sentiment orientation of the textual opinion. However, in this work, we went a step further by examining the intensity of the positive opinions, which gives a view of the level of admiration the user has for a specific entity.

(ii) The credibility and relevancy of the opinions and the users: there is a difference between an opinion that received a million likes and an opinion with no like. Same as a user who has a huge number of followers as his opinions and reviews can be seen and influence other individuals. This attribute can affect the reputation of the target entity, and can also help against the malicious fake reviews (false positive/false negative) posted by users aiming to impact the credibility and popularity of a product. Thus, our reputation system incorporates different relevancy-based features from Twitter such as the number of likes received for an opinion, number of retweets, number of user's followers expressing the opinion, and finally the user's account authenticity.

7. Conclusion

In this paper, we build the first system that generates a reputation value toward different entities (movie, product, hotel, restaurant) from user-generated data expressed on the Twitter microblogging platform. Our system incorporates three main features: sentiment orientation of the textual reviews, sentiment intensity of the positive reviews, and the credibility of the users and the tweets. The result is a numerical score between 0 and 10 that reflects the reputation of that specific entity. To evaluate the effectiveness of the proposed system, we compared its output results with the ground truth, which is the weighted average votes of thousands of users expressing their satisfaction toward the entity by giving a numerical score between 0 and 10. These weighted average votes are taken from IMDb, Amazon, TripAdvisor, and Yelp concerning respectively four products and services (movie: 7.9, phone product: 7.3, hotel: 9, restaurant: 7.1). The experimental results conducted on four real-world Twitter datasets related to the aforementioned products show that our system provides a reputation value (movie: 7.66, phone product: 7.18, hotel: 8.71, restaurant: 7.08) that is near to the ground truth. The comparison shows that our system produces a reliable reputation value that could be used in real-life applications.

Our system can serve as a decision-making tool for users and businesses in order to know the quality

of a specific product or service on Twitter. Therefore, our future studies will focus on employing other features from images and videos shared by the users on Twitter. In addition, we will attempt to improve the robustness against attacks so the system can resist attempts of entities to manipulate reputation scores. We will also expand our system to make it flexible in order to generate a reputation value from other social media platforms such as Facebook and Instagram. Finally, we will try to improve the accuracy of our sentiment classifier by employing other advanced deep learning models.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

"conceptualization, Achraf BOUMHIDI, and El Habib NFAOUI; methodology, Achraf BOUMHIDI, and El Habib NFAOUI; software, Achraf BOUMHIDI; validation, Achraf BOUMHIDI, El Habib NFAOUI; formal analysis, Achraf BOUMHIDI; investigation, El Habib NFAOUI; resources, Achraf BOUMHIDI; data curation, Achraf BOUMHIDI; writing—original draft preparation, Achraf BOUMHIDI; writing—review and editing, Achraf BOUMHIDI and El Habib NFAOUI; visualization, Achraf BOUMHIDI; supervision, El Habib NFAOUI; project administration, El Habib NFAOUI.

References

[1] P. Resnick and R. Zeckhauser, "Trust among strangers in internet transactions: Empirical analysis of eBay' s reputation system", *Advances in Applied Microeconomics*, Vol. 11, pp. 127–157, 2002.

[2] J. Schneider, G. Kortuem, J. Jager, S. Fickas, and Z. Segall, "Disseminating trust information in wearable communities", *Pers. Technol.*, Vol. 4, No. 4, pp. 245–248, 2000.

[3] Z. Yan, X. Jing, and W. Pedrycz, "Fusing and mining opinions for reputation generation", *Inf. Fusion*, Vol. 36, pp. 172–184, 2017.

[4] A. Benlahbib and E. H. Nfaoui, "A hybrid approach for generating reputation based on opinions fusion and sentiment analysis", *J. Organ. Comput. Electron. Commer.*, pp. 1–19, 2019.

[5] A. Benlahbib, A. Boumhidi, and E. H. Nfaoui, "A Logistic Regression Approach for Generating Movies Reputation Based on Mining User Reviews", In: *Proc. of International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS)*, Taza, Morocco, pp. 1–7, 2019.

[6] A. Boumhidi, A. Benlahbib, and E. H. Nfaoui, "Mining Online Opinions and Reviews Using Bi-LSTM for Reputation Generation", *Artificial Intelligence and Industrial Applications*, pp. 135–144, 2021.

[7] A. Benlahbib and E. H. Nfaoui, "Aggregating Customer Review Attributes for Online Reputation Generation", *IEEE Access*, Vol. 8, pp. 96550–96564, 2020.

[8] A. Benlahbib and E. H. Nfaoui, "MTVRep: A movie and TV show reputation system based on fine-grained sentiment and semantic analysis", *Int. J. Electr. Comput. Eng. IJECE*, Vol. 11, Art. No. 2, 2021.

[9] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis", *Min. Text Data*, pp. 415–463, 2013.

[10] A. Pak and P. Paroubek, *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*, Vol. 10. 2010.

[11] M. E. Maron, "Automatic Indexing: An Experimental Inquiry", *J. ACM*, Vol. 8, No. 3, pp. 404–417, 1961.

[12] R. Parikh and M. Movassate, "Sentiment analysis of user-generated twitter updates using various classification techniques", *Stanford, CS224N Final Report*, 118, 2009.

[13] A. L. Berger, S. A. D. Pietra, and V. J. D. Pietra, "A Maximum Entropy Approach to Natural Language Processing", *Comput. Linguist.*, Vol. 22, No. 1, pp. 39–71, 1996.

[14] G. Alec, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision", *Stanford, CS224N project report*, no. 12, 2009.

[15] C. Cortes and V. Vapnik, "Support-vector networks", *Mach. Learn.*, Vol. 20, No. 3, pp. 273–297, 1995.

[16] P. Gamallo and M. Garcia, "Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets", In: *Proc. of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, pp. 171–175, 2014.

[17] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification", *Inf. Sci.*, Vol. 181, No. 6, pp. 1138–1152, 2011.

[18] 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, 2013.

[19] X. Yao, "Evolving artificial neural networks", In: *Proc. IEEE*, Vol. 87, No. 9, pp. 1423–1447, 1999.

[20] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents", In: *Proc. of International Conference on Machine Learning*, pp. 1188–1196, 2014.

[21] R. Johnson and T. Zhang, "Effective Use of Word Order for Text Categorization with Convolutional Neural Networks", In: *Proc. of the 2015 Conf. in Computational Linguistics: Human Language Technologies*, Denver, Colorado, pp. 103–112, 2015.

[22] 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, 1980.

[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space", In: *Proc. of 1st International Conference on Learning Representations*, Arizona, USA, 2013.

[24] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation", In: *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1532–1543, 2014.

[25] U. Gupta, A. Chatterjee, R. Srikanth, and P. Agrawal, "A sentiment-and-semantics-based approach for emotion detection in textual conversations", *DBLP: journals/corr/GuptaCSA17, 1707.06996*, 2017

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[27] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", In: *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, Minneapolis, Minnesota, pp. 4171–4186, 2019.

[28] D. Araci and Z.Genc, "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models", *DBLP: journals/corr/abs-1908-10063.bib, 1908.10063*, 2019

[29] M. Peters, *et al.*, "Deep Contextualized Word Representations", In *Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, New Orleans, Louisiana, pp. 2227–2237, 2018.

[30] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification", In: *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, Melbourne, Australia, pp. 328–339, 2018.

[31] 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", In: *Proc. of the International Conference on Learning Representations*, 2020.

[32] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding", In: *Proc. of International Conference of Advances in Neural Information Processing Systems*. 2019.

[33] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", *J. Mach. Learn. Res.*, Vol. 21, No. 140, pp. 1–67, 2020.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, "Attention is all you need", In: *Proc. of the 31st International Conference on Neural Information Processing Systems*, 2017.

[35] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, "Tweeting is believing?: understanding microblog credibility perceptions", In: *Proc. of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*, Seattle, Washington, USA, 2012.