Journal of Computer Science

Original Research Paper

Unfair Reviews Detection on Amazon Reviews using Sentiment Analysis with Supervised Learning Techniques

Elshrif Ibrahim Elmurngi and Abdelouahed Gherbi

Department of Software and IT Engineering, École de Technologie Supérieure, Montreal, Canada

Article history
Received: 01-02-2018
Revised: 01-05-2018
Accepted: 11-05-2018

Corresponding Author:
Elshrif Ibrahim Elmurngi
Department of Software and IT Engineering, École de Technologie Supérieure,
Montreal, Canada
Email: elshrif.elmurngi.1@ens.etsmtl.ca

Abstract: Reputation and trust are significantly important and play a pivotal role in enabling multiple parties to establish relationships that achieve mutual benefit especially in an E-Commerce (EC) environment. There are several factors negatively affecting the sight of customers and sellers in terms of reputation. For instance, lack of credibility in providing feedback reviews, by which users might create phantom feedback reviews to support their reputation. Thus, we will feel that these reviews and ratings are unfair. In this study, we have used Sentiment Analysis (SA) which is now the subject generating the most interest in the field of text analysis. One of the major challenges confronting SA today is how to detect unfair negative reviews, unfair neutral reviews and unfair positive reviews from opinion reviews. Sentiment classification techniques are used against a dataset of consumer reviews. Precisely, we provide comparison of four supervised machine learning algorithms: Naïve Bayes (NB), Decision Tree (DT-J48), Logistic Regression (LR) and Support Vector Machine (SVM) for sentiment classification using three datasets of reviews, including Clothing, Shoes and Jewelry reviews, Baby reviews as well as Pet Supplies reviews. In order to evaluate the performance of sentiment classification, this work has implemented accuracy, precision and recall as a performance measure. Our experiments' results show that the Logistic Regression (LR) algorithm is the best classifier with the highest accuracy as compared to the other three classifiers, not merely in text classification, but in unfair reviews detection as well.

Keywords: Reputation Systems, Sentiment Analysis (SA), E-commerce (EC), Naïve Bayes (NB), Decision Tree (DT-J48), Logistic Regression (LR), Support Vector Machine (SVM)

Introduction

Nowadays, a large number of user reviews are made on almost everything that is present on the websites of the e-commerce environment, such as Amazon and eBay etc. Reviews may contain user reviews on products, destined to help other users in their buying decision making. Huge numbers of reviews exist, which makes it difficult for a consumer to read them all and make a decision. Furthermore, if the consumer reads some of the product reviews, it is difficult for them to distinguish between fair and unfair reviews. Likewise, user reviews are an important source of information for consumers. However, depending on their credibility, they can increase or decrease the reputation of products or websites.

Sentiment Analysis (SA) aims at determining the opinion of reviewers. With the growing popularity of websites such as Amazon.com where people can state their opinion on different products and rate them, e-commerce is replete with reviews and ratings. Thus, it is easy to find reviews on specific products. In this context, Reputation Systems for E-Commerce are considered as a collective measure to establish trustworthiness towards reviews or ratings coming from members of a community. Reputation systems

present a prominent technique to quantify the trustworthiness of vendors or the quality of products in E-Commerce (EC) environment. Recently e-commerce platforms, such as electronic marketplaces, have become a hot environment that allows millions of actors to trade goods and services by bringing them together. Purchasers and vendors are thereby offered incomparable opportunities to endless varieties of products. Regardless of whether Purchasers are looking for brand new technologies, highly specialized instruments or any other desired products, they will find a suitable transaction partner on the Web in most of the times. However, this "universe of strangers" also poses many issues (Dellarocas, 2005). In contrast with traditional person-to-person transactions in e-commerce, purchasers do neither get a complete feel of the products' actual quality nor do they get to know of the trustworthiness of a vendor. To tackle these issues, many e-commerce systems promote customers to provide feedback on a transaction describing their online shopping experience. Reputation systems process this information by collecting the feedback, aggregating the input data and providing one or more reputation values as output. In this way, reputation systems can assist purchasers in deciding which products or services to choose and whom to trust.

According to a recent study carried out by Diekmann *et al*. (2014),vendors with the best reputation have an increased number of sales. However, promoting trustworthy participation also bears an incentive for malicious actors to push their reputation unfairly to gain more benefit. Dishonest reviews or ratings have already become a serious problem in practice.Thus, in this research, our primary goal is detecting unfair reviews on Amazon reviews through Sentiment Analysis using supervised learning techniques in an E-Commerce environment. Our research is fundamentally focused at the document level of Sentiment Analysis, precisely on datasets of Amazon reviews. Sentiment Analysis methods will have a fundamental positive effect on reputation systems, especially inunfair reviews detection processesin an e-commerce environment and other domains. Feedback reviews in e-commerce is an important source of information for customers to reduce product uncertainty when making purchasing decisions. However, with increasing volume of feedback reviews, customers sometimes make product buying decisions based on unfair or fake feedback reviews.

One recent research provided in (Medhat *et al*., 2014) introduces a survey on different SA algorithms, however, it only concentrates on using algorithms in diverse languages, with no focus on unfair reviews detection (Kalaivani and Shunmuganathan, 2013; Singh *et al*., 2013). Detecting unfair rating and unfair reviews have been studied in several works, including (Dellarocas, 2000; Wu *et al*., 2010). The methods that are used include: Clustering ratings into unfairly lowratings and unfairly high ratings and using third-party ratings on the producers of ratings, where ratings from less reputable producers are then assumed as unfair.

This research presents four supervised machine learning algorithms that include Naïve Bayes (NB), Decision Tree (DT-J48), Logistic Regression (LR) and Support Vector Machine (SVM) in order to classify an opinion document that is put in comparison with three distinct Amazon reviews datasets. This research also spots unfair positive reviews, unfair neutral reviews and unfair negative reviews with the use of this method. The main goals of our study is to classify the document polarity of Amazon reviews datasets as fair or unfair reviews, with the use of Sentiment Analysis algorithms and supervised learning techniques.

The conducted experiments through sentiment classification algorithms have shown the performance measures of precision, recall and accuracy. In three cases (Clothing, Shoes and Jewelry reviews dataset, Baby reviews dataset and Pet Supplies dataset), we have applied NB, DT-J48, LR and SVM classifiers. These classifiers provide a useful perspective for understanding and evaluating many learning algorithms.

We can summarize the main contributions of this study as follows:

- This study use the Weka tool, an open source software for implementing machine learning algorithms (Hall *et al*., 2009), to apply sentiment classification with the NB, DT-J48, LR and SVM algorithm which classifies the Amazon reviews datasets into unfair and fair reviews
- The sentiment classification algorithms are applied with stopwords removal, using three different Amazon reviews datasets. We observed that it is more effective to use the stopwords removal method than not using stopwords and that is also more efficient to detect unfair reviews
- This work implement several analysis on various Amazon reviews datasets to getthe supervised learning algorithmswith regard to precision, recall and exactitude

The remainder of this paper is organized as per the following: Section 2 shows the related works. Section 3 presents the applied methodology. Section 4 displays the results of the experiment and lastly, Section 5 presents our conclusion and future studies.

Related Work

The majority of reputation models have been focused only on the overall products' ratings without taking into consideration their views provided by consumers (Xu *et al*., 2015). Conversely, some of the reputation models have been focused solely on the overall products' reviews without taking into consideration the ratings provided by consumers. Furthermore, most E-commerce websites let their customers add textual reviews in order to give their opinion about the product in details (Tian *et al*., 2014; Abdel-Hafez and Xu, 2013). Consumers can read these reviews and users are more and more dependent on reviews rather than on ratings. Through the Reputation, sentiment analysis methods could be used by models to extract the opinions of users and use the corresponding data in the reputation system, data that can include opinions about various features (Abdel-Hafez and Xu, 2013; Cocea *et al*., 2012).

Detection processes of sentiment classification based on a machine learning technique can clearly be expressed as a supervised learning technique with three classes: negative, neutral and positive. The testing and training data used in the existing research is commonly from reviews (Liu and Zhang, 2012).

There is fundamental importance in the identification and filtering of unfair reviews (Jindal and Liu, 2008; Moraes *et al*., 2013) proposed a method to categorize the textual review of a given topic. The document level sentiment analysis is applied for stating a positive, neutral or negative sentiment. Supervised learning algorithms consist of two stages, extraction and especially reviews' selection using supervised learning models, such as NB algorithm. However, we need the Sentiment Analysis (SA) for each class of the reviews feedback containing the product feature, in order to classify the customer feedback reviews as negative reviews, neutral reviews or positive reviews. We need also to detect unfair positive reviews, unfair neutral reviews and unfair negative reviews by using several supervised learning classification algorithms.

A major research field has emerged around the subject of how to extract the best and most accurate method and simultaneously categorize the customers' written reviews into negative or positive opinions. Such research is still in introductory preliminary phase, but much work has been done in relation to several languages (Liu and Cheng, 2005; Ku *et al*., 2006).

A survey on various applications and SA algorithms was introduced in a recent research presented in (Medhat and Korashy, 2014), however, it only concentrates on using algorithms in various languages and does not concentrate on the detections of unfair reviews (Kalaivani and Shunmuganathan, 2013; Singh *et al*., 2013).

Supervised learning is a type of machine learning that requires learning from a set of training data. However, a dataset of the product is usually represented as a corpus of documents that possesses text processing challenges to be overcome before a classification model (Shankar and Lin, 2011). Cases of text processing techniques are stopword removal and tokenization. The common classification techniques for document analysis include Support Vector Machine (Elmurngi and Gherbi, 2017), Naive Bayes (Zhang and Li, 2007), Logistic Regression (Cheng and Hüllermeier, 2009), Decision Tree (Rajput and Arora, 2013).

In this study, we present four supervised machine learning algorithms to classify the sentiment that is compared using three different Amazon reviews datasets. We also use these methods to detect unfair positive reviews and unfair negative reviews. Our study's main goal is to classify Amazon reviews datasets into fair reviews or unfair reviews with the use of Sentiment Analysis algorithms and supervised learning techniques.

The results of the conducted experiments have shown their accuracy and performance via four sentiment classification algorithms in order to detect unfair reviews. We have performed our experiments using three different datasets: The Clothing, Shoes and Jewelry reviews dataset and Baby reviews dataset. We have found that the Logistic Regression (LR) algorithm is more accurate as compared to the Naïve Bayes (NB), Support Vector Machine (SVM) and Decision Tree (DT-J48) algorithms, as much in text classification as in unfair reviews detection.

Methodology

Our methodology was organized in the next six steps, as shown in Fig. 1, steps that involve the supervised sentiment classification approaches using Weka tool for text classification as described below.

Step One: Amazon Reviews Collection

We have based our experiment on analyzing the standard dataset's sentiment value using machine learning algorithms. We have used the Amazon reviews' original dataset to test our reviews classification methods. Amazon.com has many different kinds of products, but here we would focus on three datasets: Clothing, Shoes and Jewelry reviews dataset, Baby reviews dataset and Pet Supplies dataset. The datasets are available and have been collected by (McAuley and Leskovec, 2013). Table 1 describes a summary of the three collected datasets.

Step Two: Data Cleaning

The dataset used in our experiment is obtained from Amazon product data and was divided into five scales rating: 1 star, 2 stars, 3 stars, 4 stars and 5 stars. The original dataset is not easy to model and usually not so clean. We have deleted some blank rows that cause confusion in the analysis process. The datasets before and after cleaning are listed in Table 2 and are separated to apply the sentiment classification classifiers after cleaning datasets.

Step Three: Data Preprocessing

Data preprocessing is a significant step in the text mining process and plays an important part in a number of supervised learning techniques. We have broken down data preprocessing as per the following:

StringToWordVector (STWV)

StringToWordVector filter is the main text analysis tool in Weka and it makes the transformed datasets' attribute value either Positive, Negative or Neutral for all single-words, depending on the word appearing in the document or not. It's a filtration process which is used by the following two sub-processes: Stopwords Removal and Tokenization.

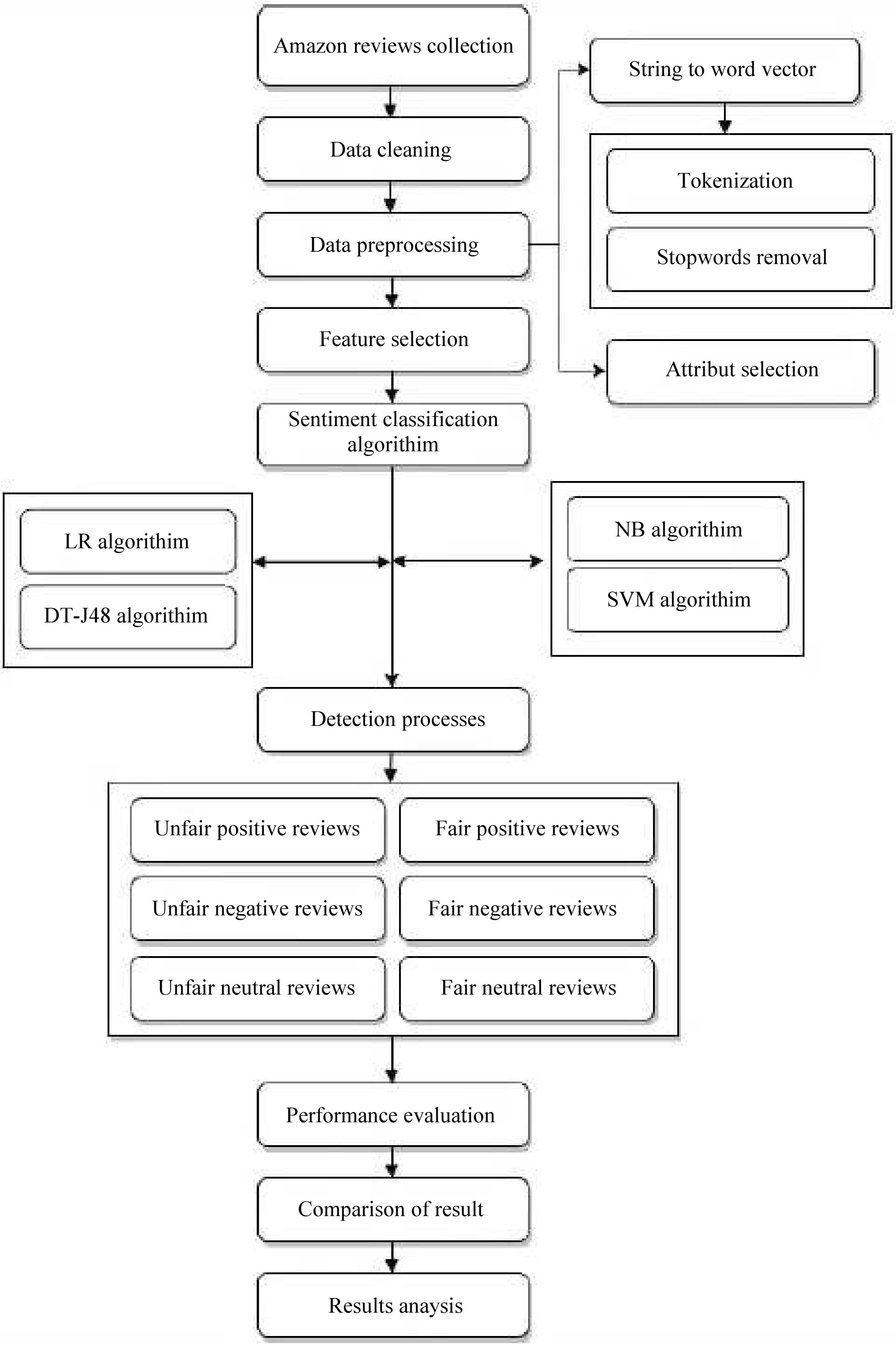

Fig. 1: Steps used in the supervised learning approach

Table 1: Number of reviews and ratings of dataset

Dataset	Reviews	Ratings
Clothing, Shoes and Jewelry	278,677	278,677 (1to5 scores)
Baby	160,792	160,792 (1to5 scores)
Pet Supplies	157,836	157,836 (1to5 scores)

Table 2: Datasets before and after cleaning

	Before cleaning			After cleaning		
Dataset	View of a dataset	Class rating	Number of reviews	View of a dataset	Class rating	Number of reviews
Clothing, shoes and jewelry	ReviewerID, asin (ID of the product), reviewerName, helpful (rating of the review), reviewText, overall (rating of the product), summary (summary of the review), unixReviewTime, reviewTime	1 star, 2 star 3 star 4 star,5 star	26655 30425 221597	ReviewText, overall (rating of the product)	Negative Neutral Positive	23019 30423 221578
Baby	ReviewerID, asin (ID of the product), reviewerName, helpful (rating of the review), reviewText, overall (rating of the product), summary (summary of the review), unixReviewTime, reviewTime	1 star,2 star 3 star 4 star,5 star	17012 17255 126525	ReviewText, overall (rating of the product)	Negative Neutral Positive	17001 17252 126479
Pet supplies	reviewerID, asin (ID of the product), reviewerName, helpful (rating of the review), reviewText, overall (rating of the product), summary (summary of the review), unixReviewTime, reviewTime	1 star,2 star 3 star 4 star,5 star	17655 15933 124248	ReviewText, overall (rating of the product)	Negative Neutral Positive	12314 8106 118203

Stopwords Removal and Tokenization

Stopwords are common words that must be filtered out, before training the classifier. Some of those words are common words (e.g., "the," "a," "I," "of," "you," "and," "it") but do not add any significant information to our labeling scheme and do not add value to a sentence's meaning, but instead they bring confusion to our classifier.

Attribute Selection

Attribute selection in machine learning, also known as feature selection, is the process of selecting a subset of relevant features for use in model construction. Attributes selection can significantly increase the classification accuracy and make it better.

Step Four: Feature Selection

Feature Selection (FS) methods in sentiment analysis have got a significant role in increasing classification accuracy and identifying relevant attributes (Koncz and Paralic, 2011). Our research has implemented one feature selection method (BestFirst + CfsSubsetEval, GeneticSearch) largely used for the SA classification task with Stopwords Removal. Our analysis of Amazon reviews datasets with feature selection method found the use of Logistic Regression (LR) algorithm gave more accuracy in the classification task.

Step Five: Sentiment Classification Algorithms

For this step, the Sentiment classification algorithm was used to classify documents as positive, negative, or neutral. In our study, we used four popular supervised classifiers such as NB, DT-J48, LR and SVM classifiers.

Naïve Bayes(NB)

In machine learning Techniques, The NB algorithm is based on the Bayes rule of conditional probability with independence assumptions between the features.

Decision Tree (DT-J48)

The DT is a predictive machine-learning technique that decides the target value of a new sample based on several attribute values of the available data. DT-J48 is the implementation of Ross Quinlan's Iterative Dichotomiser 3 algorithm, used to generate a decision tree from a dataset.

Logistic Regression (LR)

The LR is a classification algorithm, also called the logistic function, used to assign observations to a discrete set of classes. logistic regression is actually a robust technique for two-class and multiclass classification. It is a simple, fast and popular classification technique. In our study, we used this algorithm and found it to be the best and most accurate method.

Support Vector Machine (SVM)

The SVM is supervised learning techniques with related learning algorithms that analyze dataset used for classification. In recent years, the SVM has been among the most widely used and most popular classifiers with supervised learning techniques.

Step Six: Detection Processes

This step consists in predicting the models output on testing the datasets and then generating a confusion matrix that classifies the reviews into positive, negative or neutral ones. The following attributes are involved in the results:

- True Positive Reviews (TPR): Fair Positive Reviews found in the testing data and defined as the number of sentences that are correctly predicted by the classification model as Positive
- False Positive Reviews (FPR): Unfair Positive Reviews found in the testing data and defined as the number of sentences that are incorrectly predicted by the classification model as Positive
- True Negative Reviews (TNR): Fair Negative Reviews found in the testing data and defined as the number of sentences that are correctly predicted by the classification model as Negative
- False Negative Reviews (FNR): Unfair Negative Reviews found in the testing data and defined as the number of sentences that are incorrectly predicted by the classification model as Negative
- True Neutral Reviews (TNR): Fair Neutral Reviews found in the testing data and defined as the number of sentences that are correctly predicted by the classification model as Neutral
- False Neutral Reviews (FNR): Unfair Neutral Reviews found in the testing data and defined as the number of sentences that are incorrectly predicted by the classification model as Neutral

In Table 3, the confusion matrix shows the number of fair and unfair predictions made by the model compared with the actual classifications, equations 1 to 9 displays numerical parameters that could be applied following measures to evaluate the performance of detection process. For each algorithm used in this study, there is a different confusion matrix and evaluation of performance.

The confusion matrix represents a particularly significant part of our research since it lets us classify the Amazon datasets reviews into unfair or fair reviews. The confusion matrix is applied to each of the two algorithms mentioned in Step 4.

Step Six: Comparison of Results

Here, we compared the different accuracy and precision provided by the Amazon reviews datasets using different classification algorithms and identified which algorithm was the most significant in the detection of Unfair positive and negative and Neutral Reviews.

Experimentsand Result Analysis

In this section, we present our experimental results from four different supervised machine learning algorithms to classify sentiment of three datasets, which are Clothing, Shoes and Jewelry reviews dataset, Baby reviews dataset and Pet Supplies dataset. Moreover and at the same time, we have used the same approaches to detect unfair reviews using Weka 3.8 tool, which is the latest stable version.

Confusion Matrix

Using the confusion matrix is one of the approaches used to evaluate the performance of a classifier. For a given set of a classifier and a document, there are six possible outcomes: True negative, false negative, true neutral and false neutral, true positive and false positive. If the document is labelled negative and is classified as negative, then it is counted as fair negative, else, if it is classified as positive then it is counted unfair positive. Likewise, if a document is labelled positive and is classified as positive, then it is counted as fair positive, else, if it is classified as negative, then it is calculated as unfair negative. Similarly, if a document is labelled neutral and is classified as neutral, then it is calculated as fair neutral, else, if it is classified as negative or positive, then it is calculated as unfair negative or positive.

The confusion matrix displays the number of fair and unfair predictions acquired from the classification model in comparison with the actual results. The confusion matrix is obtained by implementing NB, DT-J48, LR, SVM algorithms.

Table 4, 5 and 6 display confusion matrix for the Clothing, Shoes and Jewelry reviews dataset and the Baby reviews dataset, respectively.

Table 3: The confusion matrix

		Predicted class A Fair	Predicted class B Unfair	Predicted class C Unfair
Actual class A	Fair	True Negative Reviews (TNR)	False Neutral Reviews (FNeR)	False Positive Reviews (FPR)
Actual class B	Unfair	False Negative Reviews (FNR)	True Neutral Reviews (TNeR)	False Positive Reviews (FPR)
Actual class C	Unfair	False Negative Reviews (FNR)	False Neutral Reviews (FNeR)	True Positive Reviews (TPR)

$$\text{Unfair Negative Reviews Rate} = \text{FNR}/\text{TNR} + \text{FNeR} + \text{FPR} \tag{1}$$

$$\text{Unfair Neutral Reviews Rate} = \text{FNeR}/\text{FNR} + \text{TNeR} + \text{FPR} \tag{2}$$

$$\text{Unfair Positive Reviews Rate} = \text{FPR}/\text{FNR} + \text{FNeR} + \text{TPR} \tag{3}$$

$$\text{Fair Negative Reviews Rate} = \text{TNR}/\text{TNR} + \text{FNeR} + \text{FPR} \tag{4}$$

$$\text{Fair Neutral Reviews Rate} = \text{TNeR}/\text{TNeR} + \text{FPR} + \text{FNR} \tag{5}$$

$$\text{Fair Positive Reviews Rate} = \text{TPR}/\text{TPR} + \text{FNeR}+\text{FNR} \tag{6}$$

$$\begin{aligned}\text{Accuracy} = {} & \text{TPR} + \text{TNR} + \text{TNeR}/\text{TNR} + \text{FNRclassB} + \text{FNRclassC} + \text{FNeR} \\ & + \text{TNeR} + \text{FNeR} + \text{FPRclaasA} + \text{FPRclassB} + \text{TPR}\end{aligned} \tag{7}$$

$$\text{Precision} = \text{TNR}/\text{TNR} + \text{FNR class B} + \text{FNRclass C} \tag{8}$$

$$\text{Recall} = \text{TNR}/\text{TNR} + \text{TNeR} + \text{FPR} \tag{9}$$

Table 4: Confusion matrix on clothing, shoes and jewelry

			Predicted class A	Predicted class B	Predicted class C
Algorithms			Fair	Unfair	Unfair
NB	Actual class A	Fair	6304	2118	14597
	Actual class B	Unfair	3551	3794	23078
	Actual class C	Unfair	2118	5387	211621
DT-J48	Actual class A	Fair	5183	1310	16526
	Actual class B	Unfair	2713	2248	25462
	Actual class C	Unfair	2979	2008	216591
LR	Actual class A	Fair	5006	1129	16884
	Actual class B	Unfair	2354	2151	25918
	Actual class C	Unfair	2470	1806	217302
SVM	Actual class A	Fair	2835	86	20098
	Actual class B	Unfair	1386	84	28953
	Actual class C	Unfair	1879	101	219598

Table 5: Confusion matrix on baby reviews dataset

			Predicted class A	Predicted class B	Predicted class C
Algorithms			Fair	Unfair	Unfair
NB	Actual class A	Fair	353	3253	1267
	Actual class B	Unfair	172	14234	7030
	Actual class C	Unfair	46	6707	15925
DT-J48	Actual class A	Fair	322	3479	1072
	Actual class B	Unfair	237	14800	6399
	Actual class C	Unfair	94	7545	15039
LR	Actual class A	Fair	380	3427	1066
	Actual class B	Unfair	199	15131	6106
	Actual class C	Unfair	50	7610	15018
SVM	Actual class A	Fair	303	3610	960
	Actual class B	Unfair	188	15633	5615
	Actual class C	Unfair	122	8179	14377

Table 6: Confusion matrix on pet supplies dataset

			Predicted class A	Predicted class B	Predicted class C
Algorithms			Fair	Unfair	Unfair
NB	Actual class A	Fair	5436	919	5959
	Actual class B	Unfair	2341	1059	4706
	Actual class C	Unfair	2956	1141	19857
DT-J48	Actual class A	Fair	5554	523	6237
	Actual class B	Unfair	2275	534	5297
	Actual class C	Unfair	2829	541	20584
LR	Actual class A	Fair	5220	438	6656
	Actual class B	Unfair	2094	513	5499
	Actual class C	Unfair	2317	426	21211
SVM	Actual class A	Fair	4150	252	7912
	Actual class B	Unfair	1537	308	6261
	Actual class C	Unfair	1683	196	22075

Evaluation Parameters

For us to establish the performance evaluation of the four Classification algorithms, we use an experiment on three different product reviews in terms of Unfair Negative Reviews predictive value, Unfair Neutral Reviews predictive value, Unfair Positive Reviews predictive value, Fair Negative Reviews predictive value, Fair Neutral Reviews predictive value, Fair Positive Reviews predictive value. Table 7, 8 and 9 display the evaluation parameters' results for four different classifiers and provide a summary of the experiment's recordings.

The graph in Fig. 2, 3 and 4 show a rate of Unfair Negative Reviews predictive value, Unfair Neutral Reviews predictive value, Unfair Positive Reviews predictive value, Fair Negative Reviews predictive value, Fair Neutral Reviews predictive value and Fair Positive Reviews predictive value from the comparative analysis of four different algorithms.

Classifier Evaluation Metrics: Accuracy and Precision and Recall for Various Datasets

Table 10 displays the results of evaluation parameters for four different Classification algorithms, including: NB, DT-J48, LR, SVM algorithms and provides a summary of this experiment's results.

Table 7: Evaluation parameters on clothing, shoes and jewelry

Algorithms	Unfair negative reviews %	Unfair neutral reviews %	Unfair positive reviews %	Fair negative reviews %	Fair neutral reviews %	Fair positive reviews %
NB	3.2	3.1	70.5	27.4	12.5	95.5
DT-J48	2.3	1.4	78.6	22.5	7.4	97.7
LR	1.9	1.2	80.1	21.7	7.1	98.1
SVM	1.3	0.1	91.8	12.3	0.3	99.1

Table 8: Evaluation parameters on baby reviews dataset

Algorithms	Unfair negative reviews %	Unfair neutral reviews %	Unfair positive reviews %	Fair negative reviews %	Fair neutral reviews %	Fair positive reviews %
NB	3.1	2.1	74.8	27.2	7.6	96.3
DT-J48	2.5	0.6	81.3	24.1	2.8	98.0
LR	2.2	0.8	80.6	24.1	3.9	98.0
SVM	2.5	0.1	86.0	20.6	0.4	98.1

Table 9: Evaluation parameters on pet supplies dataset

Algorithms	Unfair negative reviews %	Unfair neutral reviews %	Unfair positive reviews %	Fair negative reviews %	Fair neutral reviews %	Fair positive reviews %
NB	16.5	5.7	52.2	44.1	13.1	82.9
DT-J48	15.9	2.9	56.5	45.1	6.6	85.9
LR	13.8	2.4	59.5	42.4	6.3	88.5
SVM	10	1.2	69.4	33.7	3.8	92.2

Table 10: Comparison of accuracy, precision, recall and time taken to the build model (in seconds) of classifiers on baby reviews dataset

		Evaluation metrics %		
Algorithms	Class	Precision	Recall	Time taken to the build model (seconds)
NB	neg	43.7	27.4	17.71
	neu	33.6	12.5	
	pos	84.9	95.5	
DT-J48	neg	47.7	22.5	261.55
	neu	40.4	7.4	
	pos	83.8	97.7	
Logistic Regression	neg	50.9	217.0	83.81
	neu	42.3	7.1	
	pos	83.5	98.1	
SVM	neg	46.5	123.0	34122.09
	neu	31.0	0.3	
	pos	81.7	99.1	

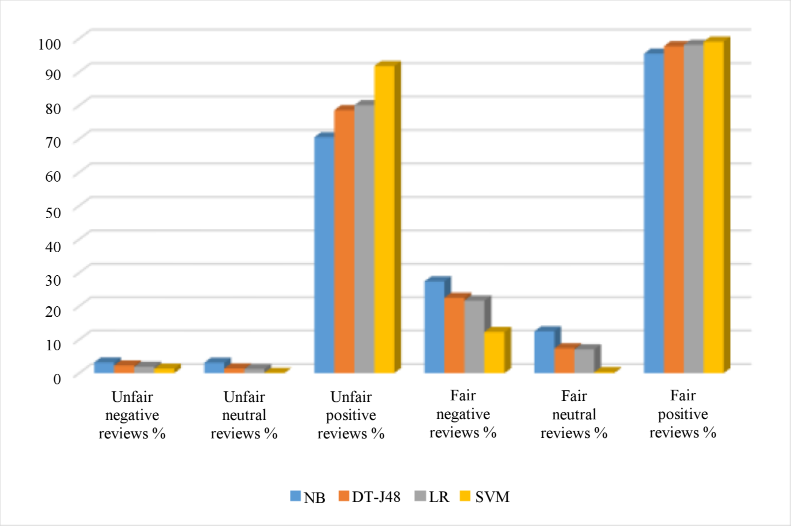

Fig. 2: Graph showing the evaluation parameters on clothing, shoes and jewelry dataset

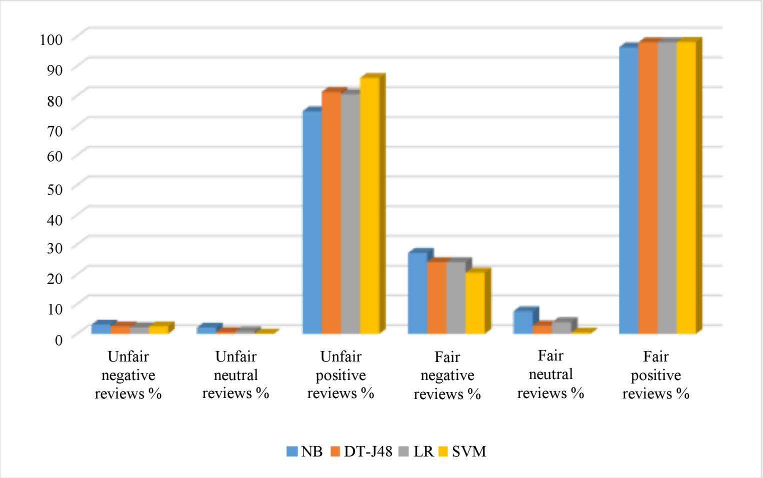

Fig. 3: Graph showing the evaluation parameters on baby reviews dataset

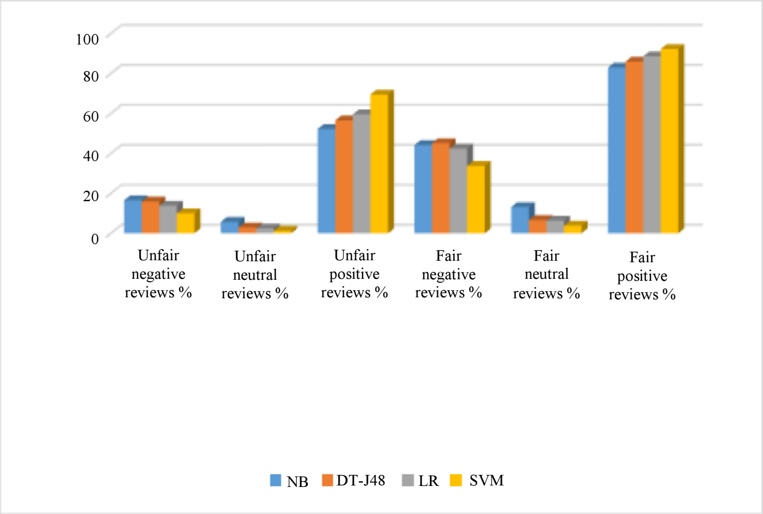

Fig. 4: Graph showing the evaluation parameters on pet supplies dataset

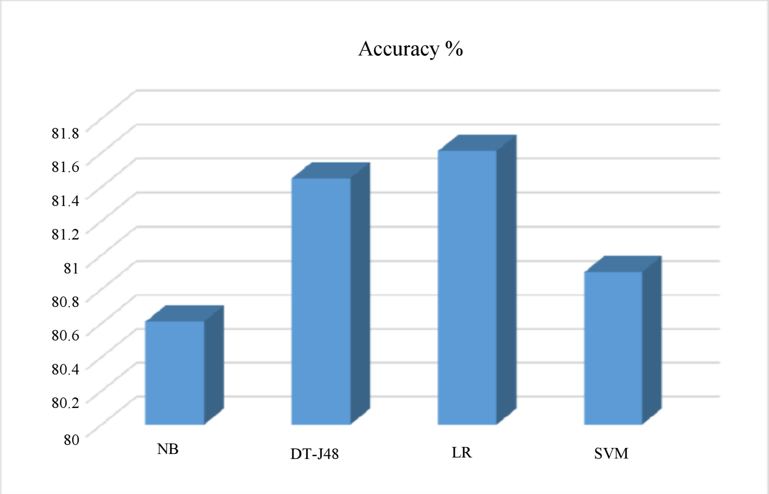

Fig. 5: Comparison of accuracy of different classifiers on clothing, shoes and jewelry reviews dataset

The Comparison of accuracy of different classifiers on Clothing, Shoes and Jewelry reviews dataset in Table 11 indicates that the LR algorithm outperformed NB, DT-J48, SVM algorithms.

The graph in Fig. 5 displays Accuracy of evaluation parameters for NB, DT-J48, Logistic Regression, SVM algorithms, as applied on the Musical Instruments reviews dataset. The Logistic Regression algorithms classification accuracy outperformed other algorithms.

The Comparison of accuracy of different classifiers on Baby reviews dataset and Clothing, Shoes and Jewelry reviews dataset and pet supplies reviews dataset in Table 12, 13 and 14 indicate that the LR algorithm outperformed NB, DT-J48, SVM algorithms.

The graph shown in Fig. 6 displays Accuracy of evaluation parameters for NB, DT-J48, LR, SVM algorithms, as applied on the Baby reviews dataset. The Logistic Regression algorithm's classification accuracy outperformed other algorithms.

Table 11: Classification Accuracy of different algorithms

Algorithms	Accuracy %
NB	80.61
DT-J48	81.45
LR	**81.61**
SVM	80.90

Table 12: Comparison of accuracy, precision, recall and time taken to the build model (in seconds) of classifiers on baby reviews dataset

		Evaluation metrics %		
Algorithms	Class	Precision	Recall	Time taken to the build model (seconds)
NB	neg	51.0	27.2	10.45
	neu	30.8	7.6	
	pos	82.6	96.3	
DT-J48	neg	53.6	24.1	97.05
	neu	36.6	2.8	
	pos	81.7	98.0	
LR	neg	56.1	24.1	67.78
	neu	36.2	3.9	
	pos	81.8	98.0	
SVM	neg	49.6	20.6	11561.03
	neu	34.5	0.4	
	pos	80.8	98.1	

Table 13: Classification accuracy of different algorithms

Algorithms	Accuracy %
NB	79.45
DT-J48	79.94
LR	**80.09**
SVM	79.37

Table 14: Comparison of accuracy, precision, recall and time taken to the build model (in seconds) of classifiers on pet supplies dataset

		Evaluation metrics %		
Algorithms	Class	Precision	Recall	Time taken to the build model (seconds)
NB	neg	50.6	44.1	1.94
	neu	34.0	13.1	
	pos	65.1	82.9	
DT-J48	neg	52.1	45.1	18.89
	neu	33.4	6.6	
	pos	64.1	85.9	
Logistic Regression	neg	54.2	42.4	12.34
	neu	37.3	6.3	
	pos	63.6	88.5	
SVM	neg	56.3	33.7	16085.65
	neu	40.7	3.8	
	pos	60.9	92.2	

Table 15: Classification Accuracy of different algorithms

Algorithms	Accuracy %
NB	59.38
DT-J48	60.10
LR	**60.72**
SVM	59.79

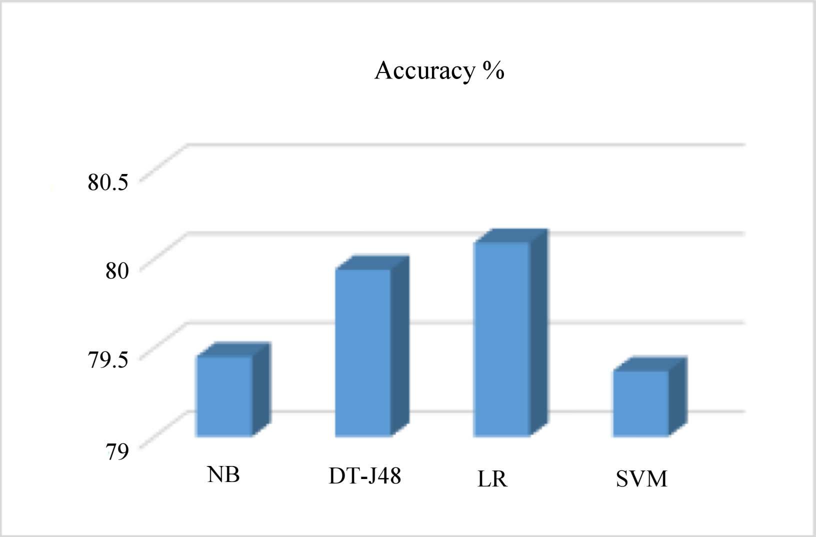

Fig. 6: Comparison of accuracy of different classifiers on baby reviews dataset

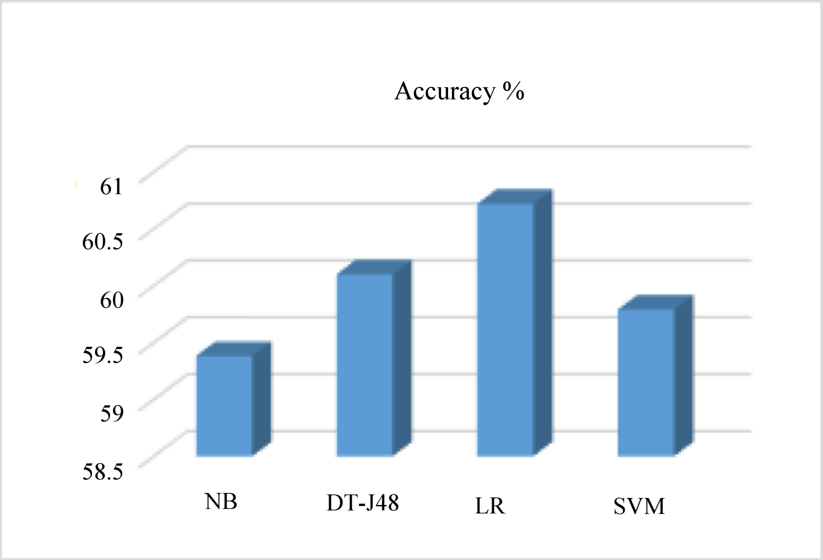

Fig. 7: Comparison of accuracy of different classifiers on pet supplies reviews dataset

The Comparison of accuracy of different classifiers on Baby reviews dataset in Table 15 indicates that the LR algorithm outperformed NB, DT-J48, SVM algorithms.

The graph shown in Fig. 7 displays Accuracy of evaluation parameters for NB, DT-J48, LR, SVM algorithms, as applied on the pet supplies reviews dataset. The Logistic Regression algorithm's classification accuracy outperformed other algorithms..

Discussion

Table 16 and Fig. 8 show the summary of experimental results. The experiments include four supervised machine learning algorithms, NB, DT-J48, LR, SVM algorithms to the Amazon product reviews datasets. This study could observe that well-trained supervised machine learning techniques were able to perform very useful classifications on reviews sentiment polarities (Negative, Neutral, Positive). In matters of accuracy, LR turned out to be the best algorithm for all tests, as it correctly classified 81.61% on Clothing, Shoes and Jewelry reviews dataset and 80.09% on Baby reviews dataset and 60.72% on Pet Supplies reviews dataset. Also, in our experimental results, we observed that the detection rate of unfair positive reviews is greater than the detection rate of unfair negative reviews and unfair neutral reviews.

In conclusion, from this analysis and through detecting of unfair positive reviews that the e-commerce domain is facing a problem of "all good reputation", making it difficult for purchasers to select credible sellers.

Table 16: Performance evaluation rate and accuracy for unfair reviews detection

Experiments	Unfair negative % reviews % of LR	Unfair neutral reviews % of LR	Unfair positive reviews % of LR	Accuracy of LR %
Results on clothing, shoes and jewelry reviews dataset	**1.9**	**1.2**	**80.1**	**81.61**
Results on Baby reviews dataset	**2.2**	**0.8**	**80.6**	**80.09**
Results on pet supplies reviews dataset	**13.8**	**2.4**	**59.5**	**60.72**

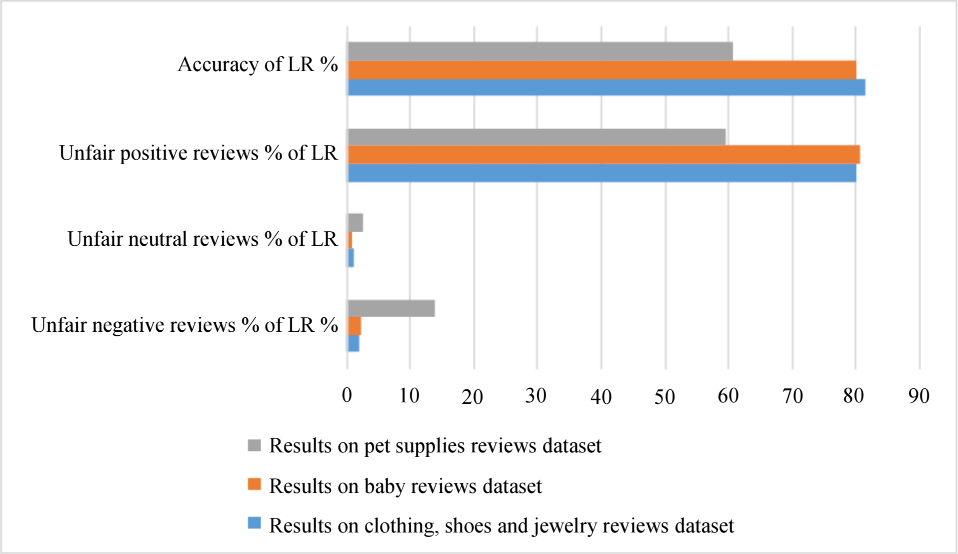

Fig. 8: Summary of experimental results

Conclusions and Future Work

In this research, we proposed NB, DT-J48, LR and SVM algorithms to analyze Amazon reviews datasets. We also presented sentiment classification methods and we carried out our experiments using three different datasets of Amazon reviews with stopwords removal.

Our experimental approaches studied the accuracy, precision and recall of sentiment classification algorithms. Moreover, we were able to detect unfair negative reviews, unfair neutral reviews and unfair positive reviews using the detection processes of this method.

The main contributions of this study are summarized as follows:

- Firstly, this study compares different sentiment classification algorithms in Weka tool, which are used to classify Amazon reviews datasets into fair and unfair reviews
- Secondly, this study implements one feature selection method used for the SA classification task and tests with Stopwords Removal to find the best-supervised learning algorithm in terms of accuracy

For future work, we wish to extend this work to use more recent snapshot Amazon reviews datasets as well as different feature selection methods. Additionally, we may use sentiment classification methods to detect unfair reviews and unfair ratings using different tools, such as Statistical Analysis System (SAS) or software machine learning library (scikit-learn) and then we would evaluate our work performance using these tools.

Acknowledgement

This study was supported and funded by the Libyan Ministry of Education and Canadian Bureau for International Education (CBIE).

Author’s Contributions

Elshrif Elmurngi: Writing the manuscript and implementing and organizing the research plan.

Abdelouahed Gherbi: Research project supervision.

Ethics

This article is original contribution of the authors. There are no ethical issues included in this article.

References

Abdel-Hafez, A. and Y. Xu, 2013. A survey of user modelling in social media websites. Comput. Inform. Sci., 6: 59-71. DOI: 10.5539/cis.v6n4p59

Cheng, W. and E. Hüllermeier, 2009. Combining instance-based learning and logistic regression for multilabel classification combining instance-based learning and logistic regression for multilabel classification weiwei cheng and eyke hüllermeier. Machine Learn., 76: 211-225. DOI: 10.1007/s10994-009-5127-5

Cocea, M., S. Weibelzahl, E. Menasalvas and C. Labbe, 2012. SDAD 2012 The 1st international workshop on sentiment discovery from affective data. Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (KDD' 12), Bristol, UK.

Dellarocas, C., 2000. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. Proceedings of the 2nd Conference on Electronic Commerce, Oct. 17-20, ACM, Minneapolis, Minnesota, USA, pp: 150-157. DOI: 10.1145/352871.352889

Dellarocas, C., 2005. Reputation mechanism design in online trading environments with pure moral hazard. Inform. Syst., 16: 209-230. DOI: 10.1287/isre.1050.0054

Diekmann, A., B. Jann, W. Przepiorka and S. Wehrli, 2014. Reputation formation and the evolution of cooperation in anonymous online markets. Am. Sociol. Rev., 79: 65-85. DOI: 10.1177/0003122413512316

Elmurngi, E. and A. Gherbi, 2017. An empirical study on detecting fake reviews using machine learning techniques. Proceedings of the 7th International Conference on Innovative Computing Technology, Aug. 16-18, IEEE Xplore Press, Luton, UK, pp: 107-114. DOI: 10.1109/INTECH.2017.8102442

Hall, M., E. Frank, G. Holmes, B. Pfathringer and P. Reutemann *et al*., 2009. The Weka data mining software: An update. SIGKDD Explorat., 11: 378-382. DOI: 10.1145/1656274.1656278

Jindal, N. and B. Liu, 2008. Opinion spam and analysis. Proceedings of the International Conference on Web Search and Data Mining, Feb. 11-12, ACM, Palo Alto, California, USA. DOI: 10.1145/1341531.1341560

Kalaivani, P. and K. Shunmuganathan, 2013. Sentiment classification of movie reviews by supervised machine learning approaches. Ind. J. Comput. Sci., 4: 285-292.

Koncz, P. and J. Paralic, 2011. An approach to feature selection for sentiment analysis. Proceedings of the 15th International Conference on Intelligent Engineering Systems, Jun. 23-25, IEEE Xplore Press, Poprad, Slovakia, pp: 357-362. DOI: 10.1109/INES.2011.5954773

Ku, L., Y. Liang, H. Chen, K. Lun-Wei and L. Yu-Ting *et al*., 2006. Opinion extraction, summarization and tracking in news and blog corpora. Artificial Intelligence.

Liu, B., M. Hu and J. Cheng, 2005. Opinion observer: Analyzing and comparing opinions on the Web. Proceedings of the 14th International Conference on World Wide Web, May 10-14, ACM, Chiba, Japan. DOI: 10.1145/1060745.1060797

Liu, B.B. and L. Zhang, 2012. A Survey of Opinion Mining and Sentiment Analysis. In: Mining Text Data, Aggarwal, C.C. and C. Zhai (Eds.), Springer US, Boston, MA, ISBN-10: 978-1-4614-3223-4, pp: 415-463.

McAuley, J. and J. Leskovec, 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. Proceedings of the 7th Conference on Recommender Systems, Oct. 12-16, ACM, Hong Kong, China , pp: 165-172. DOI: 10.1145/2507157.2507163

Medhat, W.W., A. Hassan and H. Korashy, 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams Eng. J., 5: 1093-1113. DOI: 10.1016/j.asej.2014.04.011

Moraes, R., J.F. Valiati and W.P.G. Neto, 2013. Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Syst. Applic., 40: 621-633. DOI: 10.1016/j.eswa.2012.07.059

Rajput, S. and A. Arora, 2013. Designing spam model-classification analysis using decision trees. Int. J. Comput. Applic., 75: 975-8887. DOI: 10.5120/13145-0549

Shankar, S. and I. Lin, 2011. Applying machine learning to product categorization. Department of Computer Science, Stanford University.

Singh, V.K., R. Piryani, A. Uddin and P. Waila, 2013. Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. Proceedings of the International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing, Mar. 22-23, IEEE Xplore Press, Kottayam, India, pp: 712-717. DOI: 10.1109/iMac4s.2013.6526500

Tian, N., Y. Xu, Y. Li, A. Abdel-Hafez and A. Jøsang, 2014. Product feature taxonomy learning based on user reviews. Webist.

Wu, G., D. Greene, B. Smyth and P. Cunningham, 2010. Distortion as a validation criterion in the identification of suspicious reviews. Proceedings of the 1st Workshop on Social Media Analytics, Jul. 25-28, ACM, Washington D.C., District of Columbia, pp: 10-13. DOI: 10.1145/1964858.1964860

Xu, G., Y. Cao, Y. Zhang, G. Zhang and X. Li *et al*., 2015. TRM: Computing reputation score by mining reviews.

Zhang, H. and D. Li, 2007. Naïve bayes text classifier. Proceedings of the International Conference on Granular Computing, Nov. 2-4, IEEE Xplore Press, Fremont, CA, USA, pp: 708-711. DOI: 10.1109/GrC.2007.40