Fake Reviews Identification Based on Deep Computational Linguistic Features

Saleh Nagi Alsubari[1], Mahesh B. Shelke[2], Sachin N. Deshmukh[3]
Department of Computer Science & Information Technology,
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad,
salehalsubri2018@gmail.com
mahesh_shelke21@hotmail.com
sndeshmukh@hotmail.com

Abstract

E-commerce platform has become an important resource of information. It takes into account the feedbacks of consumers about products and services purchased from the online website, these feedbacks are named as reviews. Online websites provide consumers with the ability to write product or service reviews after buying, so that when new customers make decisions to buy products or services from the online website, they read the recommendations or reviews written by people who have experienced the product or service. Those reviews, however, may be trusted (real) or spam (fake) reviews. E-commerce website fraudsters who deceive potential customers and reputation businesses or defame them can intentionally write fake reviews. Consequently, fake review detection techniques are essentially required for classification of reviews as fake (spam) or trusted (genuine) review. Main objective of this paper is to analyze, identify and detect the fake reviews of electronic products dataset that relate to different USA cities. In this paper, we investigate several feature extraction techniques such as LIWC, sentiment analysis, POS and subjectivity. Based on these methods, we extract set of features from the review text like authenticity, analytic thinking, polarity, objective, subjective, counts of adjective, verb, nouns and adverbs. For feature selection, we used an IG (Information Gain) to select discriminative and highest features. Three different supervised machine-learning techniques are Decision tree, Random forest and Adaptive boosting are applied for classification the reviews as fake or trusted and the achieved results were 96 %, 94% and 97 % in the term of accuracy respectively.

Keywords: *Fake reviews, Fraudsters, Fake review detection, feature extraction, feature selection and E-commerce platform*

1. Introduction

An advance in web 2.0 has increased the movement towards online purchases via Ecommerce Website. Internet access is increasingly growing nowadays due to its availability in both rural and urban areas making the world digital. Most of consumers procure their daily needs such as products, or services from online Ecommerce websites, so before purchasing process takes place, they go through posted reviews to see the experience of previous consumers towards products or services. Fake reviews posted in Ecommerce websites represent opinions of customers in which these reviews play a crucial role in e-business because they can indirectly affect future buying decisions. Manufacturing companies are currently using customer reviews to detect product problems and find information about their competitors on market intelligence. As these reviews effect the buyer’s side, several persons provide deceptive reviews to improve the purchasing of products found on sites of e-commerce. These people are primarily known as review spammers and their practices are called as review spamming. Review spamming involves adding misleading or false information in reviews to misguiding customers and affecting company revenues. Fake opinions can be classified into three types: 1) Untruthful (fake) opinions. 2) Review on brand only.3) Non-reviews. Untruthful (fake) opinions can be written deliberately to deceive readers or opinion mining systems. Such reviews represent unworthy positive reviews (opinions) for particular target products in order to support them and give negative reviews to worthy products for defaming them. This type of review is known as hyper spam review. Second type of fake opinions is review on brand only; these reviews can be posted and affected the brand of suppliers or retailers. Third type of fake opinions is non-reviews which consists of two subsets such as (a) Announcements and (b) unrelated reviews, both include inquiries, replies or unspecified texts [1]. Large numbers of positive reviews encourage a customer to buy product and enhance manufacturer's financial gain, while negative reviews lead customers to seek alternatives and thus cause financial losses [2]. Since customer's reviews can have a major impact on the credibility of the brands and products, so companies will be encouraged to generate positive deceptive reviews to their own brands and deceptive reviews on their competitors' brands [3]. There are different ways to spam the online e-commerce website with deceptive reviews for example, hiring specialist firms specialized in generating spam reviews, employing crowd-sourcing sites to use review spammers or using automated feedback software bots [4]. Reviews posted by those who have not experienced the topics are known as fake reviews and the person who produces the fake reviews is named as an individual review spammer [5, 6]

2. Literature Survey

Online reviews are widely used by individual consumers to make purchasing decisions online, i.e. whether or not to buy a particular product and by manufacturing companies to detect product problems and to find information about their competitors on market intelligence. Fake review detection is a popular research topic in the last two decades. Many re-searchers have performed several studies on spam review due to its significant effect on e-commerce websites.

The first research work for the reliability of the reviews was by [1]. With respect to our literature survey, it presented spam review analysis study based on amazon review dataset. It implemented a logistic regression for classification the reviews into spam and nospam. The result performance of logistic regression technique was 78 % in the term accuracy. In paper [7], the authors have employed a similar method in which the correlation between pairs of reviews that are calculated using a probabilistic language model. The authors calculated the probability of similarity between two reviews by using Kullback Leibler divergence metric, which determines the difference between two probability distributions.

For investigating the linguistic differences between both truthful and spam reviews, so the authors of this study observed that spam reviews that focus on the information given on a product page are more difficult to be read than true reviews[11].

In reference [8], authors have introduced research for detecting review spammers using behavioral features. They developed a model to classify spammers based on amazon's product reviews dataset (11,083 labeled reviews and reviewers) by using linear regression approach.

Analyzing the results of yelp filtering fake reviews algorithm used in yelp.com website. This algorithm is employed to filter fake and truthful reviews. The used dataset in this study is real-life yelp dataset that consists of 5678 reviews and 5124 hotel reviewers in addition to 58517 reviews and 35593 restaurant reviewers. Two types of features studied in this experiment, which are linguistic features that include word unigram, word bigram and part of speech. Regarding reviewers' behavioral features that consist of a higher number of reviews, review length, proportion of positive reviews, maximum content and similarity reviewers' deviation. The accuracy reported was 86% with implementation of SVM technique [9].

In paper [10], authors have familiarized a similar language-based technique, which concentrates on semantic analysis with FrameNet that helps to comprehend the features of spam reviews as compared to true reviews. The authors include two methods of statistical analysis (normalized frame rate and standardized bi-frame rate) to analyze the semantic structures of hotel reviews and identify semantic differences in spam and nonspam reviews.

Based on four lexical characteristics [15], authors have studied the difference between an authentic and fictitious through three online hotels. The dataset used in this experiment consist of 1900 hotel reviews that collected from different domains websites that were Trip advisor.com, Hotels.com and United.com. These characteristics include comprehensibility, specificity, exaggeration and negligence. For analyzing purpose, they have used logistic regression approach.

In reference [8], authors have suggested framework for fake reviews features, these authors have used in their work yelp products reviews dataset and Random forest and Ada boost techniques for classification of reviews and reviewers based on behavioral and text content centric features. The results gained from this experimentation were 82 % in the term of f1-score for both classifiers.

In order to discover spam reviews from online hotel reviews based on stylometric and lexical features, authors in this works [13] have used two different classification techniques that are SMO (Sequential Minimal Optimization) and Naïve Bays techniques for detecting spam reviews. The achieved results were 81% and 70 % in terms of F1-score performance metric respectively.

3. Methodology for System Development

The followed methodology in this research work is shown in Figure (1). Different steps are explaining the methodology for fake reviews detection .It begins with a dataset that collected from the Yelp e-commerce website by means of Web scraping. Then, features are well defined and calculated to train the classifiers for detecting fake reviews.

3.1. Scraping process:

In general term, web scraping can be defined as the extracting and crawling data from a website by some tools. The first step in the above-cited methodology focuses on the data collection. For our experiment, the dataset has been crawled and collected from the yelp website, which has a particular algorithm to filter the reviews into fake or truthful reviews, and highly reportedly accurate in [10]. Furthermore, it is utilized as a reference for labeling the dataset used in [15].

3.2. Preprocessing steps:

Second phase of the proposed methodology is to perform the data cleaning which includes the below steps.

3.2.1. NA values Removal: NA values affect the performance of the classifiers, if there are such values in the dataset, the classification process will not be done so that we remove all NA values from rows and columns of the dataset.

3.2.2. One-word review removal: In this step, we have deleted a review which has only single word in dataset.

3.2.3. Extra spaces removal: as large space between review contents makes some problems for the next phase of the model development, so we drop them from the whole dataset.

3.3. Feature Extraction Methods:

In this phase, four feature extraction methods have been implemented for extracting important and helpful list of words as feature set to be input for the classifiers in order to detect a review as fake or trust. These methods are LIWC, POS, Sentiment and Subjectivity.

3.3.1. LIWC stands for Linguistic Inquiry and word count as an analysis tool, which can be used to analyze, extract and calculate significant feature from contents of the texts. It offers 90 output variables by using this method, so we investigate two measures for review text and reviewer as follow:

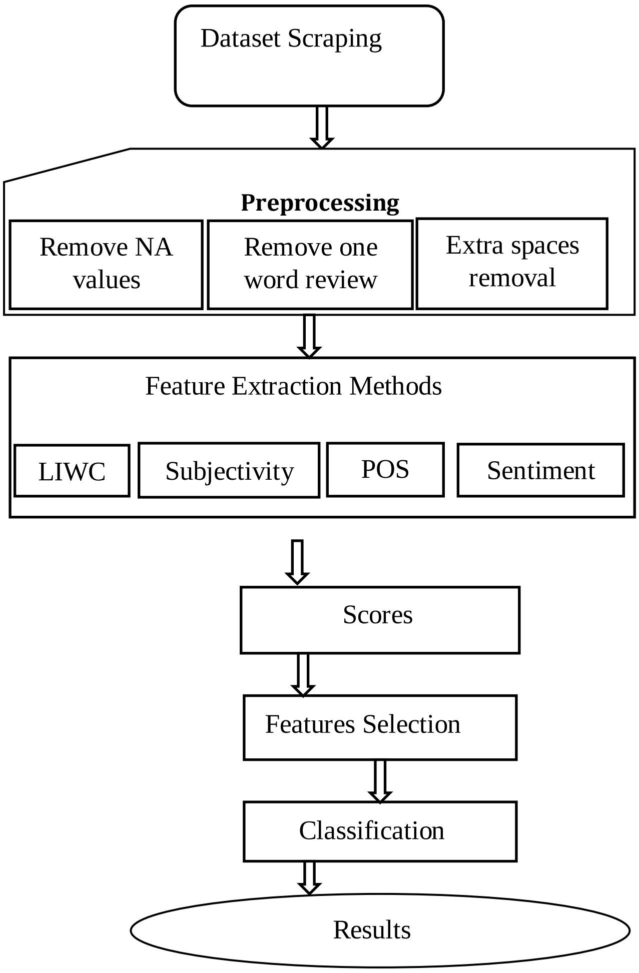

Fig1: graphical representation of Methodology

3.3.1.1. Authenticity indicates personalized and honest writing of the reviewers. Once people authentically or honestly expose themselves in their writing, they become more intimate, modest, and vulnerable [14]. Every review text of the whole dataset has been analyzed and calculated by following authenticity equation.

$$A(r) = \sum FPS + TPS + TPP + excl\ (differ) - Negemo - Motion \qquad (1).$$

Where A(r) refers to authenticity of the review. FPS, TPS and TPP are representing a frequency and total percentage of First Person Singular, Third Person Singular and Third Person Plural pronouns as well as percentage of Negative, exclamation, differ, Motion words in the text review. We have performed computation and analysis of all review texts of the dataset based on LIWC authenticity variable that has the output value in range 1 to 100 degree, so we have concluded that the trusted reviews have greater than or equal to 49 % score, whereas the fake ones have less than 48.5 % authenticity score.

3.3.1.2. Analytical Thinking:

Analytical thinking is one of LIWC variables based on eight-function words. It is considered to analyze and capture the degree of the individuals use words that indicate logical, formal and thinking patterns [15]. People of low analytic thought have a tendency to write and communicate in a concise and narrative language, whereas those have high analytic thinking generally give a better performance in writing the professional language. For identifying the analytic thinking of the reviewer (person who writes fake or trusted review about product or service in online e-commerce website), we have employed the below formula.

$$AT = \sum 30 + \text{Articles} + \text{Prep} - \text{PP} - \text{IMP} - \text{auxvrb} - \text{conj} - \text{adver-Negation} \qquad (2)$$

Where 30 value was added to words percentage in order to make the output of the equation always positive.

Articles, Prep, PP, IMP, auxverb, conj, adverb and Negation represent the total percentage of Articles, Preposition, Personal Pronouns, Impersonal pronouns in addition auxiliary verbs, conjunction, adverbs and negation words in the text review. After completing the process of calculating an analytic thinking of reviewers based on their texts reviews, we have found out that fake reviews have greater than 75 % analytic score while the trusted ones have less than 75 %. This clearly specify spammer reviewers (people who write and produce spam /fake reviews in order to fame or defame product or service in online e-commerce website) have higher analytic thinking than non-spammer reviewers (consumers who write truthful reviews after buying a product or service from online e-website).

3.3.2. POS tagging:

POS tagging is one of the feature extraction methods used in text classification. It can be well defined as the process of attachment of each word in the review text with part of speech tag based on locating its context in the sentences of the review. After implementing this method and counting of all parts of speech in the review text, we have inferred that the trusted reviews have more nouns and, adjectives whereas fake ones have more verbs and adverbs.

3.3.3. Sentiment analysis:

Sentiment analysis includes the study of an analysis of the text, the processing of natural languages, computational linguistics to recognize, extract and analyze the subjective knowledge from textual data. It is used in this an experiment to calculate the polarity of the text review and find positive, neutral as well as negative reviews by using below mentioned formula.

$$S(r) = \frac{P(W) - N(W)}{T(W)} \qquad (3)$$

Where S(r) indicates a sentiment (S) of review, P (W) is related to the number of positive words in the text review, N (W) is indicates the number of negative words in the same the text review. T (W) is indicates the total number of positive and negative words in the text review. The above formula produces one of three values that are 1 (positive review), -1 (negative review) and 0 (neutral review).After implemented the sentiment analysis formula, we inferred that fake reviews posted by spammers contain strong positive (for fame a product or service) and strong negative words in order to defame product or service.

3.3.4 Subjectivity: Purpose of subjectivity is to extract significant features from the review text. It uses for computing and identifying the review text as subjective or objective. Based on the following formula, we have calculated the subjectivity scores for every review in datasets.

$$Sub = 1 - P(w) + N(w) \qquad (4)$$

Where P (w) and N (w) refer to the number of positive and negative words in the text. The subjectivity formula produces two values that are zero and one. The objective review is specified by zero value .It is based on facts and called an unbiased review. Furthermore, a subjective review indicated by one value and based on personal experience, it named as a biased review.

3.4. Scores Generation:

This step calculates scores values for different linguistic features of every single review of entire dataset. Specific numbers of values can represent these scores. Every feature has particular range as can be defined and shown in a table below.

Table 1: Summaries of scores for every feature

Feature name	Range
Authenticity Score	1 To 100
Analytic thinking Score	30 To 100
Polarity score	-1 To 1
Subjective score	0 To 1
Nouns count	1 To 80
Adjectives count	1 To 30
Verbs count	1 To 40
Adverbs count	0 To 10
Rating value	1 To 5
Word count	2 To 360

3.5. Features Selection:

The process of selecting a discriminative subset of related features is a known as feature selection. Fundamental premise while using a selection approach is the data that comprises some features, which are either repetitive or insignificant and thus they omitted with next to no information losing. In this research work, we study information gain as feature selection to select highest information features in order to be kept and inputted to machine learning algorithm for detecting of fake and trusted reviews.

3.5.1. Information Gain

Information Gain [16] is one of feature selection techniques used in a text classification, which calculates the textual information gained after understanding the value of the feature in the text. It performs the calculation based on entropy that is used to measure the instability of the probability distribution of every textual feature in the given classes. We have decided and labelled the scores of authenticity analytic thinking features based on threshold values that have lower and upper limits. In case of authenticity and analytic thinking scores the lower (weak) and upper (strong) limits represented by less or greater than 50 and 75 % respectively. According to all others features scores were decided based on the output of their formulas. Table 2 describes the labeling of features scores of the review text of dataset .After completing the labeling process of all features then Information Gain and entropy are implemented.

Table 2: Description and labeling of features scores

Feature name	Range	Labelled Score
Authenticity Score	1 To 99	If (score <= 50 %)? Weak: Strong
Analytic thinking	30 To 99	If (score < 75 %)? Weak: Strong

ISSN: 2005-4238 IJAST

Score		
Polarity score	-1 To 1	If (score < 0)? Negative: Positive
Subjective score	0 To 1	If (score = 0 ‖ score =1)? Objective: Subjective
Nouns count	0 To 100	If (count <=10)? less: More
Adjectives count	0 To 30	If (count <= 10)? Less: More
Verbs count	0 To 40	If (count <= 15)?Less: More
Adverbs count	0 To 10	If (count < 5)? Less: More
Rating value	1 To 5	If (value <= 3)? Low: High
Word count	2 To 360	If (count <= 136)? Short: Long

The formulas for both entropy and information gain are demonstrated in the following section.

$$H(C) = \sum - P_1(C_i)\log_2(C_i) - P_2(C_i)\log_2(C_i) \qquad (5)$$

Where P (C_i) points out to the probability of how many reviews texts belong to the trusted and fake classes. If feature set X has *n* of different values X = $\{x_1, x_2, x_3 \dots x_n\}$, then an entropy is calculated for X feature as follow:

$$H(C|X) = \sum - P(C_i|x_i)\log_2(C_i|x_j) - P(C_i|x_j)\log_2(C_i|x_j) \qquad (6)$$

For fake review detection, we generally categories the reviews into fake (spam / deceptive) and trusted (truthful/ non spam) reviews. After extraction, calculating the values of all features of a review text and entropy is applied on these features, the next step is information gain (IG) to compute and select the feature that has the highest information to be a root node in the classifier.

$$IG(X) = H(C) - H(C|X) \qquad (7)$$

Depend on above formulas, we have calculated information gain for all review text features as shown in the below figure 2.

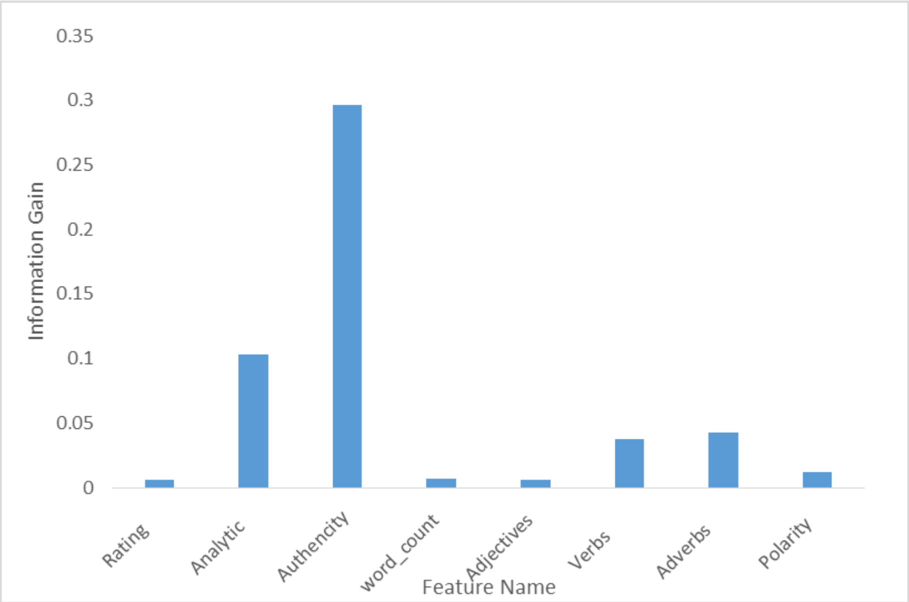

Fig-2: Graphical presentation of calculation of information gain for review text features.

ISSN: 2005-4238 IJAST

3.6. Classification:

Before the classification process starts, we divided the dataset into 80 % as train set and 20% as test set. In order to classify and detect the review text as fake or trusted, three different supervised machine-learning algorithms based tree were applied. These algorithms are Random forest, Decision tree and Adaptive boost.

4. Experimental Results and Accuracy Measurements

4.1. Dataset Description

The dataset contains 30476 reviews of electronics product devices collected from the yelp website. An exploratory analysis of dataset reveals that it is compounded from four different USA cities as described in the below table 3:

Table 3.The size of reviews per USA city

City Name	Fake Reviews	Trusted reviews
Los Angeles	**6270**	**6009**
Miami	**1696**	**1767**
NY	**3865**	**3979**
San Francisco	**3642**	**3248**

4.2. Accuracy Measurements

We have trained the classifier with 24380 samples and tested the 6096 samples for classification. As table 3 shows the results for false classified for RF, DT and Ada boost are 158,153, 73.200,86 and 125 And also True Classified are 2951, 2956,3036, 2787,2901 and 2862 in both classes respectively. This clarifies that RF (Random Forest) has more misclassified samples than other classifiers.

Table 4. Summaries of confusion matrixes for the used classifiers.

Name of the classifier	True Fake	False Fake	True Trusted	False Trusted
RF	2951	158	2787	200
DT	2956	153	2901	86
Ada boost	3036	73	2862	125

Accuracy= (TT+TF) / (FF+FT+TT+TF) (8)

Precision (Trusted) =TT/ (TT+FT) (9)

Precision (fake) =TF/ (TF+FF) (10)

Recall (fake) =TF/ (TF+FT) (11)

Recall (Trusted) = TT/ (TT+FF) (12)

F1-score (fake) = 2 * (precision (fake) * Recall (fake)) / (precision (fake) * Recall (fake)) (13)

F1-score(Trusted)=2*(precision(Trusted)*Recall(Trusted)) / (precision(Trusted)

*Recall (Trusted)) (14)

For analyzing the classification performance of classifiers, we have employed different evaluation parameters along with their equation as demonstrated above. These have included the accuracy, recall, precision and f1-score. While calculation of accuracy, we have observed that, the adaptive boosting classifier is performed better than other classifiers in the term of parameters evaluations. The below table reveals the results of classification performance of all used classifiers.

Table 5. The classification performance for the used classifiers.

Name of the classifier	Class Name	Precision %	Recall %	F1-score %	Accuracy %
RF	Fake	0.94	0.94	0.93	0.94
	Trusted	0.93	0.95	0.94	
DT	Fake	0.95	0.97	0.96	0.96
	Trusted	0.97	0.95	0.96	
Ada boost	Fake	0.98	0.96	0.97	0.97
	Trusted	0.96	0.97	0.96	

Table 6. Comparative analysis of existing algorithms.

References	Dataset used	Features used	Algorithm	Result
[1]	Amazon’s product reviews dataset	Features of the review, reviewer and product	LR	78%
[9]	1600 hotel reviews dataset through AMT and Trip advisor.com	Review’s features	NB SVM	89 % 93 %
[13]	1600 hotel reviews dataset from Mechanical Turk and Trip Advisor Amazon dataset;’	Review’s features	SMO NB	81 % 71 %
[18]	Yelp dataset	Behavioral features of reviewer	Back propagation neural network	95 %
[17]	1900 hotel reviews	Review’s features	C 4.5 RF	69 % 78 %
[12]	Yelp reviews dataset	Features of reviewer	Ada Boost RF DT	82 % 81% 80 %
Our Work	Yelp reviews dataset	Review’s features	RF DT Ada boost	94 % 96 % 97 %

5. Conclusions

Due to the current possible effect of fake reviews on consumer behavior and decision purchasing, fake review detection has gained significant attention in both academic research and business domains. In this research work, we tackled the problem of fake reviews in the customer electronics domain based on deep linguistics features that relate to centric reviews features. As shown in this work, the identification of fake reviews by reading them is a challenging task for humans. Therefore, considering a set of textual features for the purpose of detecting and differentiating between fake and trusted reviews are essential. we have selected subset related features by applying IG (Information Gain) that uses to calculate a piece of information contained in every feature and the results show that authenticity followed by analytic thinking features have the highest information than other features. There are differences between the fake and trusted reviews based on authenticity and analytic thinking variables of LIWC that represented by scores produced by these variables. After analyzing the obtained scores, we found that trusted reviews have greater than 49 % and less than 75% in authenticity and analytic thinking scores respectively, whereas the fake ones have less than 48.5 % and greater than 75% respectively. Furthermore, fake reviews have either strong positive and negative sentiment. For the classification performance, we have observed that adaptive boost outperforms other classifiers in the term of accuracy. Another conclusion of this paper, according to the literature survey of fake review detection, there is no large scale labelled dataset for training the classifier. In next article, we will attempt to consider the review and reviewers' centric features in order to detect a fake reviews in an on line e-commerce websites.

6. Acknowledgments

The author would like to thank to Data Analytics Research Lab, Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad for providing infrastructure to carry out my research work.

References

[1] Jindal, Nitin, and Bing Liu, "Opinion spam and analysis." Proceedings of the 2008 international conference on web search and data mining, (2008), pp. 219-230.

[2] Vidanagama, Dushyanthi U. Thushari P. Silva, and Asoka S. Karunananda, "Deceptive consumer review detection: a survey." Artificial Intelligence Review 53.2, (2020), 1323-1352.

[3] Rayana, Shebuti, and Leman Akoglu, "Collective opinion spam detection: Bridging review networks and metadata." Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining, (2015), pp. 985-994.

[4] Wijnhoven, Fons, and Anna Theres Pieper. "Review spam criteria for enhancing a review spam detector." (2018).

[5] Heydari, Atefeh, "Detection of review spam: A survey." Expert Systems with Applications 42.7, (2015), 3634-3642.

[6] Mukherjee, Arjun, "What yelp fake review filter might be doing?" Seventh international AAAI conference on weblogs and social media, (2013).

[7] Lai, C. L, "High-order concept associations mining and inferential language modeling for online review spam detection." 2010 IEEE International Conference on Data Mining Workshops. IEEE, 2010, pp. 1120-1127.

[8] Ong, Toan, Michael Mannino, and Dawn Gregg, "Linguistic characteristics of shill reviews." Electronic Commerce Research and Applications 13.2 (2014), pp 69-78.

[9] Ott, Myle. "Finding deceptive opinion spam by any stretch of the imagination," Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1. Association for Computational Linguistics, (2011), pp. 309-319.

[10] Kim, S., Chang, H., Lee, S., Yu, M., & Kang, J, "Deep semantic frame-based deceptive opinion spam analysis." Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, (2015), pp. 1131-1140).

[11] Banerjee, Snehasish, and Alton YK Chua, "Theorizing the textual differences between authentic and fictitious reviews." Internet Research (2017).

[12] Barbado, Rodrigo, Oscar Araque, and Carlos A. Iglesias, "A framework for fake review detection in online consumer electronics retailers." Information Processing & Management 56.4, (2019), pp. 1234-1244..

[13] Dewang, Rupesh Kumar, and A. K. Singh, "Identification of fake reviews using new set of lexical and syntactic features." Proceedings of the Sixth International Conference on Computer and Communication Technology (2015), pp. 115-119.

[14] Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. Lying words: Predicting deception from linguistic styles." Personality and social psychology bulletin 29.5 (2003), pp. 665-675.

[15] Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. "The development and psychometric properties of LIWC2015", (2015).

[16] Parlar, Tuba, and Selma Ayşe Özel. "A new feature selection method for sentiment analysis of Turkish reviews." 2016 International Symposium on INnovations in Intelligent SysTems and Applications (INISTA). IEEE, (2016),pp. 1-6.

[17] Banerjee, S., Chua, A. Y., & Kim, J. J. Using supervised learning to classify authentic and fake online reviews. In Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication, (2015), pp. 1-7.

[18] Goswami, K., Park, Y., & Song, C. "Impact of reviewer social interaction on online consumer review fraud detection." Journal of Big Data, 4(1), (2017), pp 1-19.