Detection of Fake Product Reviews Using BI Directional Long Short –term Memory and One-dimensional Convolution Neural Network

[1]OJO, Adebola K. and [2]AKO, Ayito

Department of Computer Science, University of Ibadan, Nigeria
adebola_ojo@yahoo.co.uk, ayito.ako@gmail.com

Abstract

This study presents an approach to extracting data from amazon dataset and performing some preprocessing on it by combining the techniques of Bi-Directional Long Short-Term Memory and 1-Dimensional Convolution Neural Network to classify the opinions into targets. After parsing the dataset and identifying desired information, we did some data gathering and preprocessing tasks. The feature selection technique was developed to extract structural features which refer to the content of the review (Parts of Speech Tagging) along with extraction of behavioral features which refer to the meta-data of the review. Both behavioral and structural features of reviews and their targets were extracted. Based on extracted features, a vector was created for each entity which consists of those features. In evaluation phase, these feature vectors were used as inputs of classifier to identify whether they were fake or non-fake entities. It could be seen that the proposed solution has over 90% of the predictions when compared with other work which had 77%. This increase was as a result of the combination of the bidirectional long short-term memory and the convolutional neural network algorithms.

Keywords: *Fake reviews detection, Opinion Spam, Behavioral features, Convolution Neural Network, Bi-Directional Long Short-Term Memory*

1. Introduction

The Internet and the World Wide Web have spread dramatically over the past years, resulting in the creation of a variety of tools to empower people. Through the accessibility provided by modern web technologies any individual has the ability to express their opinion and share information with one another. Around 40% of the world population has an Internet connection today and web pages, forums and blogs quickly became home to an abundance of user generated content, i.e. content created exclusively by web users. Current e-commerce statistics state that 40% of worldwide Internet users have bought products or goods online via desktop, mobile, tablet or other online devices. [1]

The Web has dramatically changed the way that people express themselves and interact with others. They can now post reviews of products at merchant sites (e.g., amazon.com) and express their views in blogs and forums. It is now well recognized that such user generated contents on the Web provide valuable information that can be exploited for many applications. It is now quite common for people to read reviews on the Web for many purposes. For example, if one wants to buy a product, one typically goes to a merchant site (e.g., amazon.com) to read some reviews of existing users of the product. If the reviews are mostly positive, one is very likely to buy the product. If the reviews are mostly negative, one will most likely buy a different product. Positive opinions can result in significant financial gains and/or fames for organizations and individuals. This gives good incentives for review/opinion spam [2].

Online reviews are increasingly used by individuals and organizations to make purchase and business decisions. Positive reviews can render significant financial gains and fame for businesses and individuals. Unfortunately, this gives strong incentives for imposters to game the system by posting fake reviews to promote or to discredit some target products or businesses. Such individuals are called *opinion spammers* and their activities are called *opinion spamming*. In the past few years, the problem of spam or fake reviews has become widespread, and many high-profile cases have been reported in the news [3] [4] [5].

Figure 1: Product review examples from Amazon. [16]

Consumer sites have even put together many clues for people to manually spot fake reviews [6]. There have also been media investigations where fake reviewers admit to have been paid to write fake reviews [7]. Many businesses have tuned into paying positive reviews with cash, coupons, and promotions to increase sales. There has been a lot of speculation and evidence about the prevalence of deceptive product reviews, i.e., fake customer reviews that are written to sound authentic in order to promote the business [8].

"What other people think" has always been an important piece of information for most of us during the decision-making process. In terms of sellers and service providers, reviews usually comment on quality of service experienced, and dependability or trustworthiness of the provider. This affects our decision to buy a product or pay for a service.

Businesses and corporations display great determination in recreating large amounts of fake reviews, either positive or negative, according to their pursued goal, i.e. promote their products or demote those of their competitors. Fake reviews come in many shapes and forms, and rely on either review text, rating (or all of the above) in order to deceive unsuspecting users. Researchers have carried out some research works in this domain and have turned up solutions and models which have been tried with varying degrees of success. One of such detection strategies is the one proposed by [9].

They proposed the use of a network-based model by applying iterative multi-level algorithm to detect reviews, using Amazon dataset. The problem with this method is that the weighted values used to determine fake reviews can be repeated from an already computed weight which may give a less accurate result.

2. Related Works

Li F. *et al*. [9] proposed semi-supervised learning for spam detection in both review and reviewer levels. They designed a two-view type of semi-supervised model to cover both labeled and unlabeled data [10].

Lim E. P. *et al*. [10] focused to extract user centric features and behaviors instead of reviews' characteristics. In this study, they concentrated on review pattern and rating to discover a variety of spamming behaviors without considering review contents and its opinion polarity.

Lu Y. *et al*. [11] proposed a novel algorithm which could implement spam review and spammer detection at the same time. In this study, they used extracted features from both reviews and reviewers, and exploited review factor graph (RFG) structure to combine all these features and demonstrate belief propagation between reviews and reviewers. The main problem of this approach is the indicators which they used to label reviews as spam, such as a positive / negative trend, or high rating of reviews, as usually they cannot be good indicators, e.g., usually people express their satisfaction of product or services through positive reviews without mentioning any negative points and assign high rating.

In this study, a deep neural network [12] sequence model was applied, which was a bidirectional long short-term memory with conditional random fields (Bi-LSTM-CRF), to extract target expressions in opinionated sentences. Based on the targets extracted, sentences were classified into three groups- non-target, one-target and multi-target. Afterwards, one-dimensional convolutional neural networks (1d-CNNs) were trained for fake review sentiment classification on each group separately. Finally, the fake review sentiment polarity of each input sentence was predicted by one of the three 1d-CNNs.

3. Methodology

This section discusses the methodology involved in this study. This is shown in Figure 2.

The proposed solution is majorly divided into three main parts which are:

a. Data collection,
b. Sentiment Analysis, and
c. Visualization and results.

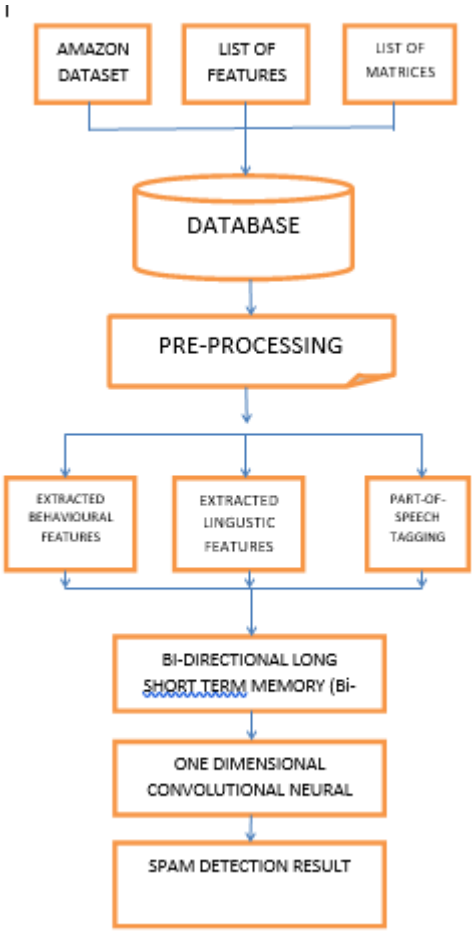

Figure 2: The Proposed Model

3.1 Data Collection

In the first step, data was collected from cloud media or social media by extracting reviews and data regarding reviews from sources such as websites and micro-blogging platform. However, to manually copy and paste all the reviews from the considered websites into a document was time-consuming and inefficient; hence this process was automated by using a data scraping script that was developed in the work.

3.2 Sentiment Analysis

After collecting the data, the second step was to analyze them. For this purpose, the Bidirectional-Long Short-Term Memory with conditional random fields (Bi-LSTM-CRF) algorithm was used to extract keywords, entities, and the One-Dimensional Convolutional Neural Network (1d-CNN) was used to the sarcasm sentiment classification of each review that had been collected. Once the data had been analyzed, scores was then normalized and then listed with the review information which was then written to a file.

3.3 Visualization and Results

After the data analysis using the Bidirectional-Long Short-Term Memory algorithm, the next step was to present the aggregated analyzed sentiment in visual components like tables and graphs where necessary. This part of the proposed allowed researchers to compare results of this research work with other results in this problem domain.

The breakdown of the different components of the proposed solution was given below:

i. The Cloud Media,
ii. Data extraction
iii. Sentence processing
iv. Sentence analyzer
v. The Bi-LSTM-CRF and 1d-CNN module
vi. The opinion aggregator module, and
vii. Result presentation.

A detailed discussion of the various modules of the proposed solution is given in the sections that follow.

3.3.1 The Cloud Media

Software as a Service (SaaS) is cloud computing best known model and delivers applications over the Internet as subscription-based services in a pay –as- you -go model to consumer and their adoption by end users is accelerating because of ease of access, cost savings, operational efficiency and enhanced flexible business capabilities (Yasmine, 2016). Online cloud consumers use Social media platforms like Facebook, twitter, review sites, blog, discussion forum etcetera and other different web platforms to express their views and emotions on different aspects of cloud services using natural language.

3.3.2 Data Extraction

Amazon is one of the largest E-commerce sites as for that there are innumerous amount of reviews that can be seen. We used data named Amazon product data. The dataset was unlabeled and to use it in a supervised learning model we had to label the data. Three JSON files were used where the structure of the data is as follows:

"reviewerID": ID of the reviewer
"asin": ID of the product
"reviewerName": name of the reviewer Manually
"helpful": helpfulness rating of the review
"reviewText": text of the review
"overall": rating of the product
"summary": summary of the review
"reviewTime": time of the review (raw)

For data, three categories from Amazon products were selected, laptop electronics reviews, Cell Phone and Accessories Reviews and Musical Instruments product reviews which consist of approximately 48500 product reviews. Where 21600 reviews were from mobile phones, 24352 were from laptop electronics and 2548 from musical instruments data. After filtering, 37,900 of the data were used for the study. From the formats used for analyzing the review polarity we used review Text.

3.3.3 Sentence Processing and Analysis

This module was used for parsing the extracted data from the data collection. Human sentences are not easily parsed by programs, as there is substantial ambiguity in the structure of human language, whose usage is to convey meaning (or semantics) amongst a potentially unlimited range of possibilities but only some of which are germane to the particular case. So an utterance "Man bites dog" versus "Dog bites man" is definite on one detail but in another language might appear as "Man dog bites" with a reliance on the larger context to distinguish between those two possibilities, if indeed that difference was of concern. It is difficult to prepare formal rules to describe informal behavior even though it is clear that some rules are being followed.

In order to parse natural language data, we must first agree on the grammar to be used. The choice of syntax is affected by both linguistic and computational concerns; for instance, some parsing systems use lexical functional grammar, but in general, parsing for grammars of this type is known to be NP-complete. Once we have identified, extracted, and cleansed the content needed for sentiment analysis, the next step is to have an understanding of that content. In many use cases, the content with the most important information is written down in a natural language (such as English, German, Spanish, Chinese, etc.) and not conveniently tagged. To extract information from this content you will need to rely on some levels of text mining, text extraction, or possibly full-up natural language processing (NLP) techniques.

To achieve the necessary level of processing needed for the sentiment analysis preparation of the extracted reviews, the following needed to be performed:

- **Structure extraction** – identifying fields and blocks of content based on tagging. Identify and mark sentence, phrase, and paragraph boundaries – these markers are important when doing entity extraction and Natural Language Processing since they serve as useful breaks within which analysis occurs.
- **Language identification** – will detect the human language for the entire document and for each paragraph or sentence. Language detectors are critical to determine what linguistic algorithms and dictionaries to apply to the text.
- **Tokenization** – to divide up character streams into tokens. this can be used for further processing and understanding. Tokens can be words, numbers, identifiers or punctuation (depending on the use case)
- **Acronym normalization and tagging** – acronyms can be specified as "I.B.M." or "IBM" so these should be tagged and normalized.
- **Lemmatization**– reduces word variations to simpler forms that may help increase the coverage of Natural Language Processing utilities. Lemmatization uses a language dictionary to perform an accurate reduction to root words.
- **Entity extraction** – identifying and extracting entities (people, places, companies, etc.) is a necessary step to simplify downstream processing. There are several different methods:

  a. Regex extraction – good for phone numbers, ID numbers (e.g. SSN, driver's licenses, etc.), e-mail addresses, numbers, URLs, hashtags, credit card numbers, and similar entities.

  b. Dictionary extraction – uses a dictionary of token sequences and identifies when those sequences occur in the text. This is good for known entities, such as colors, units, sizes, employees, business groups, drug names, products, brands, and so on.

  c. Complex pattern-based extraction – good for people names (made of known components), business names (made of known components) and context-based extraction scenarios (e.g. extract an item based on its context) which are fairly regular in nature and when high precision is preferred over high recall.

  d. Statistical extraction – use statistical analysis to do context extraction. This is good for people's names, company names, geographic entities which are not previously known and inside of well-structured text (e.g. academic or journalistic text). Statistical extraction tends to be used when high recall is preferred over high precision.

  e. Phrase extraction – extracts sequences of tokens (phrases) that have a strong meaning which is independent of the words when treated separately. These sequences should be treated as a single unit when doing NLP. For example, "Big Data" has a strong meaning which is independent of the words "big" and "data" when used

separately. All companies have these sorts of phrases which are in common usage throughout the organization and are better treated as a unit rather than separately. Techniques to extract phrases include:

i. Part of speech tagging – identifies phrases from noun or verb clauses

ii. Statistical phrase extraction - identifies token sequences which occur more frequently than expected by chance

iii. Hybrid - uses both techniques together and tends to be the most accurate method.

3.3.4 The Bi-LSTM-CRF and 1d-CNN Module

Expressing a fake opinion is a sophisticated form of speech act widely used in online communities. In the context of sentiment analysis, it means that when one says something positive, one actually means negative, and vice versa. In this work, we proposed a different way in dealing with different sentence types so as to make it easy to extract and predict the intended sentiment expressed in the sentences. In particular, we investigated the relationship between the number of opinion targets expressed in a sentence and the fake sentiment expressed in this sentence; we proposed a framework for improving product review sentiment analysis via sentence type classification. Opinion target (hereafter, target for short) can be any entity or aspect of the entity on which an opinion has been expressed. An opinionated sentence can express sentiments without a mention of any target, or towards one target, two or more targets. We defined three types of sentences: **non-target sentences**, **one-target sentences** and **multi-target sentences**, respectively. Consider the following examples from the movie review sentence polarity dataset [13] [14].

Example 1. A masterpiece four years in the making.

Example 2. If you sometimes like to go to the movies to have fun, Wasabi is a good place to start.

Example 3. Director Kapur is a filmmaker with a real flair for epic landscapes and adventure, and this is a better film than his earlier English-language movie, the overpraised Elizabeth.

It can be seen that Example 1 is a non-target sentence. In order to infer its target, we need to know its context. Example 2 is a one-target sentence, in which the sentiment polarity of the target “Wasabi” is positive. Example 3 is a multi-target sentence, in which there are three targets: “Director Kapur”, “film” and his earlier English-language movie, the “overpraised Elizabeth”. We can observe that sentences tend to be more complex with more opinion targets, and sarcasm sentiment detection is more difficult for sentences containing more targets.

Based on this observation, we apply a deep neural network sequence model, which is a bidirectional long short-term memory with conditional random fields (Bi-LSTM-CRF), to extract target expressions in opinionated sentences. Based on the targets extracted, we classify sentences into three groups: non-target, one-target and multi-target. Then, one-dimensional convolutional neural networks (1d-CNNs) are trained for fake review or spam review sentiment classification on each group separately. Finally, the fake review sentiment polarity of each input sentence is predicted by one of the three 1d-CNNs [14].

3.3.5 One Dimensional Convolution Neural Network (1d-CNN)

The 1d-CNN takes sentences of varying lengths as input and produces fixed-length vectors as outputs. Before training, word embedding for each word in the glossary of all input sentences are generated. All the word embeddings are stacked in a matrix M. In the input

layer, embedding of words comprising current training sentence are taken from M. The maximum length of sentences that the network handles is set. Longer sentences are cut; shorter sentences are padded with zero vectors. Then, dropout regularization is used to control over-fitting.

In the convolution layer, multiple filters with different window size move on the word embedding to perform one-dimensional convolution. As the filter moves on, many sequences, which capture the syntactic and semantic features in the filtered n-gram, are generated. Many feature sequences are combined into a feature map. In the pooling layer, a max-overtime pooling operation is applied to capture the most useful local features from feature maps. Activation functions are added to incorporate element-wise non-linearity. The outputs of multiple filters are concatenated in the merge layer. After another dropout process, a fully connected *softmax* layer output the probability distribution over labels from multiple classes.

CNN is one of most commonly used connectionism model for classification. Connectionism models focus on learning from environmental stimuli and storing this information in a form of connections between neurons. The weights in a neural network are adjusted according to the training data by some learning algorithm.

That is, the greater the difference in the training data, the more difficult for the learning algorithm to adapt the training data, and the worse classification results it will produce. Dividing opinionated sentences into different types according to the number of targets expressed in them can reduce the differences of training data in each group, therefore, improve overall classification accuracy.

4. Results and Implementation

Anaconda was used as the development environment alongside Spyder which is a lunch tool for scientific analysis using Python in the Anaconda environment.

The raw data gotten from amazon were passed though pre-processing phase. The data pre-processing phase requires the following activities:

- Import Libraries (JSON, NUMPY and KERAS)
- Load Data Source (Dataset and Vocabulary)
- Process Data (tokenize, stemming, POS)

After pre-processing the data, the next step is to train and test the model. It involves the following:

- Split data into Training and Validation
- Develop the Model
- Train the Model
- Make Prediction

Figure 3 presents a sentiment predicted report

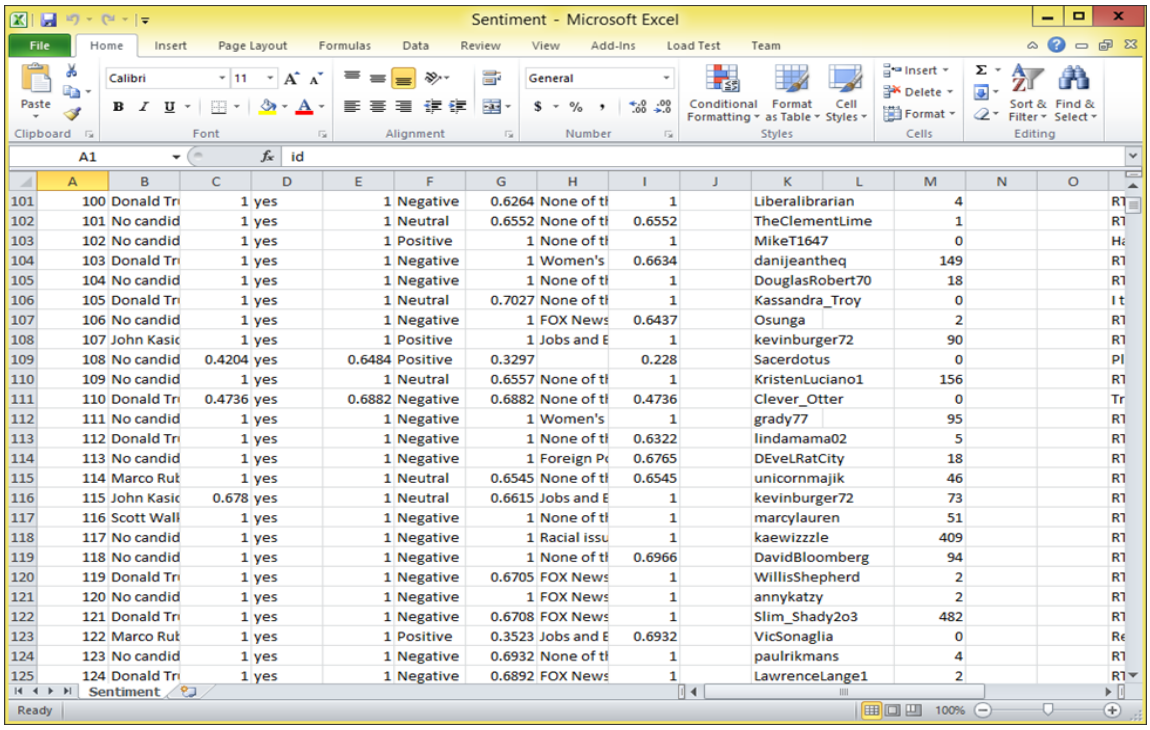

Figure 3: A sentiment predicted report

Table 1: Results of the Proposed Work

DATASET	ACCURACY
First run	91.3%
Second run	92.1%
Third run	91.5%
AVERAGE	91.6%

4.1 Comparison between the existing and the proposed work

In the existing work [15] proposed an automatic annotation method to detect spam entities by considering both behavioral and structural features along with inter- and intra-relationship among entities. They designed an algorithm to learn the network based on those features and relationship among entities. In their study, they used an iterative algorithm which, they claim, can improve the accuracy of spam detection. Since this algorithm is an iterative algorithm the weight of each edge of the network was updated iteratively. In their case, they used the variety combination of features and finding which combination gives them better and more accurate results. The result of the accuracy of the experiment is presented in table below:

Table 2: Existing Work Results

Existing Work	ACCURACY
First run	74.0%
Second run	69.0%
Third run	89.0%
AVERAGE	77.3%

Table 3: Accuracy comparison

Proposed Work	Existing Work
91.6%	77.3%

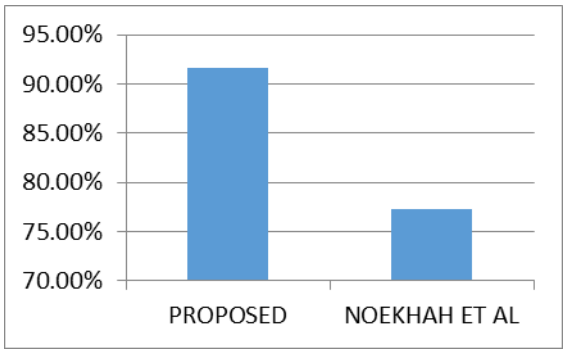

Figure 2: Accuracy comparison

It can be appreciated in Table 3 that our proposed method achieved a better accuracy result on average when compared with the existing work. It summarizes the overall reported accuracy of the proposed methodology and the existing work for the three runs of the product reviews dataset as classification classes for the evaluation. In Figure 4, it can be seen that the proposed solution over 90% of the predictions correct when compared with the existing work, which has 77%. This is a 13% increase in accuracy from the work. This increase might be as a result of the combination of the bidirectional long short-term memory and the convolutional neural network. This can be translated as a plus for the proposed model because, it means that the number of true positives made by the model surpasses that of the existing work. Our model leverages the fact that products with extensive reviewing activity include a greater amount and variety of spammer behavioral clues, as they are more likely to be heavily spammed and therefore, enable the most effective detection of fake reviews.

5. Conclusion

In this paper, we proposed a model by applying bi-directional Long Short-Term Memory (Bi-LSTM) and One-Dimensional Convolution Neural Network (1D-CNN) to detect fake product reviews in Amazon dataset. This was an improvement on the existing work. We determined a large number of behavioral and structural features to determine spam entities. In this study, we developed hybridized feature selection technique to extract structural features which refer to the content of the review (POP tagging) along with extraction of behavioral features which refer to the meta-data of the review (review's rate, reviewer id, helpfulness, etc.). We extracted both behavioral and structural features of reviews, reviewers, group of reviewers, and their targets. Based on extracted features, we created a vector for each entity which consists of those features. In evaluation phase, we also used these feature vectors as inputs of classifier to identify whether they are fake or non-fake entities. Combination of behavioral and structural features help us to improve the accuracy of our fake review detection technique which was far higher than that of the existing work.

6. Future Work

This work made a first attempt at contributing to existing literature associated with fake review detection. Due to certain constraints though, such as time limitations and the absence of annotated data, the proposed detection model was not evaluated to the extent of the

author's satisfaction. It is one of the tasks that can be reserved for future work on the field. In fact, there are already a few aspects of the problem that are open for further studying in the future. Some of these include:

i. Employing human evaluators to annotate at least a portion of an experimental dataset. This would improve the current system's detection accuracy as the impact of parameter selection could be more thoroughly evaluated.
ii. Gaining access to additional reviewing data either from Amazon or other sources. A dataset consisting of more recent records would assist in evaluating the proposed detection system on modern spam practices.
iii. Modifying the introduced methodology to better account for singleton spam reviews. While these reviews as individual pieces of content lack the influence on a product's overall rating and image, however as a whole they could pose a real threat to genuine and unsuspecting readers and consumers. Certain modifications would have to take place in regards to the spam scoring function in order to consider these spam cases.
iv. Applying the proposed fake review detection system on online sites, as a built-in functionality or an add-on/plugin.

References

[1]. Guan Wang, Sihong Xie, Bing Liu, Philip S. Yu, "Review Graph based Online Store Review Spammer Detection," in *11th IEEE International Conference on Data Mining, ICDM 2011,* , Canada, 2011.

[2]. Nitin Jindal and Bing Liu. , "Opinion Spam and Analysis," Department of Computer Science University of Illinois at Chicago 851 South Morgan Street, Chicago, IL 60607- 7053, 2008.

[3]. D. Streitfeld, "Fake Reviews, Real Problem.," 2012. [Online]. Available: http://query.nytimes.com/gst/fullpage.html?res=9903E6DA1E3CF933A2575AC0A9649D8B63. .

[4]. J. Taylor, "Are You Buying Reviews For Google Places?," 2012. [Online]. Available: http://www.localgoldmine.com/blog/reputation-management/are-you-buying-reviews-for-google-places .

[5]. DoniaGamal, Marco Alfonse, El-Sayed M. El-Horbaty, Abdel-BadeehM.Salem , "Opinion Mining for Arabic Dialects on Twitter," *Egyptian Computer Science Journal,* vol. 42, no. 4, pp. 52-61, 2018.

[6]. B. Popken, "30 Ways You Can Spot Fake Online Reviews.," 2010. [Online]. Available: http://consumerist.com/2010/04/14/how-you-spot-fake-online-reviews/. The Consumerist..

[7]. A. Kost, "Woman Paid to Post Five-Star Google Feedback," 2012. [Online]. Available: http://www.thedenverchannel.com/news/woman-paid-to-post-five-star-google-feedback.

[8]. M. Nisen, 19 September 2012. [Online]. Available: http://www.businessinsider.com/fake-reviews-are-becoming-a-huge-problem-for-businesses-2012-9 .

[9]. Li, F., Huang, M., Yang, Y., & Zhu, X., "Learning to identify review spam," in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence* , 2011.

[10]. Lim, E.P., Nguyen, V.A., Jindal, N., Liu, B., and Lauw, H.W., "Detecting product review spammers using rating behaviors.," in *19th ACM international conference on Information and knowledge management, ACM*, 2010.

[11]. Y. Lu, L. Zhang, Y. Xiao, and Y. Li., "Simultaneously detecting fake reviews and review spammers using factor graph model.," in *In Proceedings of the 5th Annual ACM Web Science Conference ACM. 2013.* , 2013.

[12]. C. R. Ivanescu, "Statistical Learning And Benchmarking: Credit Approval Using Artificial Neural Networks," *Egyptian Computer Science Journal,* vol. 43, no. 1, pp. 26-32, 2019.

[13]. Bo Pang and Lillian Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proceedings of the ACL*, 2005.

[14]. Tao Chen, Ruifeng Xu, Yulan He, Xuan Wang, "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN," *Expert Systems with Applications,* 2017.

[15]. Shirin NoekhahNaomie SalimNor Hawaniah Zakaria, "A Novel Model for Opinion Spam Detection Based on Multi-Iteration Network Structure," *Journal of Computational and Theoretical Nanoscience ,* vol. 24, no. 2, pp. 1437-1442, February 2018.

[16]. "http://blog.eckelberry.com/," 2018. [Online]. Available: blog.eckelberry.com.