International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 731-737 www.ijaem.net ISSN: 2395-5252

Fake Review Detection Using Soft Attention Based Bi-Directional LSTM

Sharad Dahate, Prof. Swati Soni
[1]*Student, Takshshila Institute of Engineering & Technology, Jabalpur, MP*
[2] *Prof, Takshshila Institute of Engineering & Technology, Jabalpur, MP*

Date of Submission: 12-01-2023 Date of Acceptance: 24-01-2023

ABSTRACT
Opinion reviews are a valuable source of information in e-commerce. Indeed, it benefits users in buying decisions and businesses to enhance their quality. However, various greedy organizations employ spammers to post biased spam reviews to gain an advantage or to degrade the reputation of a competitor. This results in the explosive growth of opinion spamming. Due to its nature and their increasing volume, fake reviews are a fast-growing serious issue on the internet. Until now, researchers have developed many Machine Learning (ML) based methods to identify opinion fake reviews. However, the traditional ML methods cannot effectively detect spam messages due to the limited feature representations and the data manipulations done by spammers to escape from the detection mechanism. As an alternative to ML-based detection, in this thesis, we proposed a Deep Learning (DL) based framework called Self Attention-based Bi-LSTM model to learn document level representation for identifying the fake reviews. Our approach computes the weightage of each word present in the sentence and identifies the spamming clues exists in the document with an attention mechanism. Then the models learns sentence representation by using Bi-directional LSTM (Bi-LSTM) as document feature vectors and identify the fake reviews with contextual information. The evaluated experiment results are compared with its variants and the result shows that proposed model outperforms other variants in terms of classification accuracy.

Index Terms— Fake Review Detection, Opcode, N-gram, Machine Learning, Fake Review Analysis, Random Forest, KNN, SVM, LSTM.

I. INTRODUCTION

People don't need to go outside to buy necessary things now. Thanks to online shops and delivery websites. Every product like foods, clothes, electronics and others can be purchased and delivered to customer's home-place through these e-commerce sites. Sales of online products hugely depend on other's opinion that already have purchased the product and used it. To gain some knowledge about a product, it is one of the best ways to watch the review section and see opinions and comments given by other users. Thus, review section has a great impact in decision making [1]. This factor also provides the opportunity to some groups of dishonest businessmen or companies to manipulate public opinion about their products. They intentionally post fake reviews for promotion of their products. Also, there are some cases where some companies are attacked by their competitors. So, researches are growing interest for fake review detection. Various models proposed to detect fake online reviews. Machine learning approaches to detect fake online reviews include: supervised classification models, semi supervised classification models and unsupervised clustering based models. Various features have been proposed by the researchers to create a better classification system. These features can be categorized into review-content based features, metadata based features, graph connectivity based features and user characteristics based features [2]. As fake user can always change their identity, content based features are most popularly used to detect deceptive online reviews. These features include word frequency count, n-grams, term frequency and inverse document frequency (TF-IDF), Parts-Of-Speech tag, noun to verb ratio and others [2]. Sentiment score as a feature is also used by many researchers. Fakes reviews are posted either to promote the products or demote it. Hence the probabilistic sentiment score is always much higher or lower when the review is deceptive.

The requirements for a more advanced online shopping system are increasing day by day. Many people prefer online shopping system now. Because online shopping system is an easier and

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 731-737 www.ijaem.net ISSN: 2395-5252

better way of shopping. That's why the authorized companies offer many types of discounts for growing business in an online platform [3]. Also, they allowed to giving customer feedback on their platform of products & services. People can post his opinion about products and its services on e-commerce sites with freedom [4]. Sharing a particular judgment about an appropriate product or its services based on their own experience is considered to be as reviews [5].

Reviews are divided into two types:

(i) Authentic reviews which are given by customer or buyer based on his personal opinion, and (ii) Reviews brought by the companies for promotion purposes [6]. These reviews are usually counted as fakes reviews [7]. Another problem in an opinion sharing websites is that spammers can quickly generate hype of the appropriate goods of spam reviews. Spam reviews play an important role in raising the value of goods or services [8]. A new person affected most of the time because when new people come on this online platform for shopping at first, he needs a judgment which good or bad product and helps to get the decision to see these product reviews [9]. For example, if a consumer wants to buy any products from the online, usually they go to the comment section and know about other buyer's feedback on this product. If the reviews are positive mostly, then the users are interested to purchase a product, otherwise they would become de-motivated to get that particular product [10]. That's why the customer gets confused about choosing his targeted product after watching these reviews and he thinking whether it will be real or fake at all.

Spamming not only causes for harmful activity but also it is using for promoting a website, especially for an e-commerce website. For increasing product rating and attracting the customer, spammers are hired for giving fake reviews of products. Identifying, removing or avoiding these kinds of spam opinion, host or authority has spent their valuable time and energy. Therefore, classifying the harmful reviews automatically becomes an urgent matter for the consumers as well as for the companies. Addressing this problem, many researchers already worked on online reviews spamming. They already proposed various techniques for preventing spam in public opinion. In this paper we have introduced an efficient technique based on some traditional machine learning algorithms for improving the accuracy of detecting spam in online reviews. Here, we prefer sentiment analyzing technique for filtering customer's opinions and their intention.

Deceptive or fake review detection has found its attention from the very beginning of this century. **Jindal et al. [11]** introduced fake review detection as a classification problem in 2008. They categorized fake reviews into deceptive and destructive reviews. They described some features to classify deceptive reviews. Following, **Ott et al. [12]** in 2011 generated a gold standard dataset for fake online detection and they used n-gram features for classification. They updated their dataset in 2013 and added more fake reviews with negative sentiment. With more data with positive and negative sentiment, the previous n-gram model's result was decreased in accuracy. Later sentiment score as a feature was used by many researchers. **Fontanarava et al. [13]** analyzed different features for fake review classification with supervised learning. They have categorized these features into two categories.

These are –

• review centric category
• Reviewer centric category.

II. LITERATURE REVIEW

Spam reviews are fictitious comments that are either machine-generated or user-generated. Both spams are challenging to identify. In recent years, with the increasing use of e-commerce online, there have been chances of fraudulent comments that play an essential role in defaming or uplifting a business. Due to the intense competition between organizations, it has become more sophisticated, and thus, many of them use the wrong approach to receive potential profit. Reviews on a product play a part in consumer decisions and build confidence in that particular product. However, they cannot be sure about the fallacy of these reviews. Spams can either be deceptive or destructive. Destructive spams are easier to identify by a typical customer since they are non-review and contain unrelated ads and messages unrelated to the product. The latter, however, may contain sentimental reviews that may be positive or negative and, thus, problematic. The existence of such reviews is crucial for the customer and the business. This concept, in other words, is also called "Opinion Mining." It is a technique in Natural Language Processing to figure out the public's mood regarding a specific product, service, or company. However, considering the deceptiveness of these reviews, these fake reviews are being used to promote a business or spread rumors and harm the reputation of competing businesses. Since the purchase decision is firmly motivated by the reviews or ratings, a study shows that work has been concluded in detecting these fraudulent reviews, but spammers' demeanor is

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 731-737 www.ijaem.net ISSN: 2395-5252

constantly developing. Spammers have been discreetly designing these fake reviews to camouflage their malevolent intentions. Many businesses appoint professionals to write inappropriate positive and negative reviews for financial gains. These are fabricated reviews that are intentionally written to seem authentic. Deceptive spam review is harmful to the repute of any product as it misleads the customer to make decisions.

Somayeh et al. [66] came up with a lexical and syntactical feature technique using machine learning classifiers to detect spam or ham. The features include n-gram, Part of speech (POS) tagging, and LIWC (Linguistic Inquiry and Word Count). They took deceptive reviews from Kaggle.com and truthful reviews from TripAdvisor.com. Their results showed 81% accuracy with Naïve Bayes (NB) classification algorithm and 70% with Sequential Minimal Optimization (SMO) using lexical features. Moreover, using syntactic features gave 76% and 69% accuracy using the same classifiers. At the same time, their combination gave 84% and 74% with NB and SMO. However, the results did not exceed 85%—furthermore;

Rajamohana et al. [67] proposed a methodology for detecting opinion spam using features detection. They proposed an approach that deals with selecting subset features from many feature sets for the classifier to separate spam or ham. The two approaches utilized are cuckoo search, and hybrid improved binary particle swarm optimization (iBPSO), Naïve Bayes, and KNN classifiers that are helping in the classification process. These two approaches have been compared, and a hybrid search achieved a comparatively higher accuracy measure. However, this approach is solely dependent on feature selection.

Moreover, **Catal and Guldan [68]** came up with supervised and unsupervised techniques to know by sight the spam review. There is a significant chance that spam reviewer is responsible for the content pollution in social media as many users have multiple login IDs. The researchers tackled that problem and utilized the most productive feature sets to structure their model. Semantic analysis is also unified in the detection process. In addition, some standard classifiers are applied on labeled datasets, and for unlabeled datasets, clustering is used after desired attributes. They worked with both labeled and unlabeled data along with a unigram model and achieved 86% results.

Ott et al. [69] proposed a model to identify fraudulent consumer reviews using multiple classifiers in online shopping. The selected classification techniques were majority voted liblinear, libsvm, minimal sequential optimization, random forest, and J48. Then the evaluation was compared with other models, SVM technique with 5-fold cross-validation to get 86% was accuracy.

Rout et al. [70] explained that how semi-supervised classifiers are used to detect online spam reviews using a dataset of hotel reviews. Dissimilar to other different kinds of spam [70] it is demanding to recognize an unreal opinion as it is needed to understand the contextual meaning to know the nature of the review. Supervised learning is conventionally used to detect fake reviews, but it also has some restrictions, such as assurance of the quality of reviews in the training dataset. Secondly, to train the classifier, it can be challenging to obtain the data because of the diverse nature of the online reviews. The limitations mentioned above can be overcome using a semi-supervised learning approach by unifying three new dimensions to the domain of the feature as in POS feature, Linguistic and Word Count Feature, and Sentimental Content features to get more significant results. A dataset of both positive and negative reviews has been used. They, however, achieved an 83% f-score.

He et al. [71] introduced the rumors model and applied the text mining technique, and extracted three notable characteristics of the content of reviews such as noun/verb ratio, important attribute word, and a specific quantifier. Trip Advisor dataset was used, and results showed that the unique vocabulary, specific quantifiers, and nouns it contains, the more valuable and truthful the review is. Moreover, the results showed 71.4% F-measure, 60% accuracy, 86% recall, and a fake evaluation value of 0.016952338. Meaning, higher the fake evaluation value, the more fake a review is. Deceptive opinions are more fictitious but sound real. People are hired by many businesses to write unjustified reviews about the products which are undistinguishable by the people.

Therefore, **Ott et al. [69]** performed a test that gave the accuracy of 57.33% of three human judges, which made this research even more valid, significant, and pithy. However, it is hard to define the semantic perspective from the data. Significant donations of the paper are; firstly, to understand the semantic better, a document level review is represented. Secondly, multiple syntax features are

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 731-737 www.ijaem.net ISSN: 2395-5252

used to make a feature combination to improve performance. Thirdly, domain-independent and domain migration experiments verify the SWNN and feature combination performance. Further, in the domain of neural networks, **Goswami et al. [72]** proposed a feature set by observing the user's social interaction behavior to recognize reviewer hoaxes. They used a neural network to analyze the feature set and compare it with other contemporary feature set in detecting spam. Features include the number of friends, followers, and number of times a user has provided enough room to form a relationship between opinion spam and social interaction behavior. Aside from neural networks, most scholars focused on supervised learning techniques.

Therefore, **Brar and Sharma [73]** proposed an approach that is used to analyze the review and reviewer-centric feature to detect fake reviews using the supervised learning technique. It provided comparatively better results than completely unsupervised learning techniques, mostly graph-based methods. A publically available large-scale and standard data set from a review site **Yelp.com [74]** has been considered here and has given more significant results. Furthermore, in the supervised learning domain, **Elmurngi and Gherbi [75]** analyzed the online reviews for movies using Sentiment Analysis (SA) methods and text classification for the sake of recognizing fake reviews. The scholars presented the classification of the movies review as positive or negative by using machine learning (ML) methods. The comparison between five individual ML classifiers, Naïve Bayes (NB), SVM, KNN- IBK, K*, and DT-J48, for sentiment analysis is made using two datasets that include movie review datasets V1.0 and movie review dataset V2.0.

Some researchers also focused on different factors in determining fake reviews, such as **Arjun Mukherjee et al. [76]** pay attention to fake reviewers groups instead of individual reviews; therefore, they came up with the frequent itemset mining method to identify the groups. Furthermore, they built a labeled dataset of the reviewers' group. The results showed that their methodology outperformed the standard classification techniques using the Kaggle dataset. In order to determine negative reviews on crowdsourcing platforms, **Parisa et al. [77]** observed the behavior of the reviews on these sites and observed the behavior of the reviews given. They indicated clues on the detection process of such manipulating reviews that are fake yet hiding in plain sight. However, this approach is risky because it relies on observations that may or may not be accurate.

III. PROPOSED WORK

In the first stage the data was gathered from the source, following that different pre-processing step was undertaken such as eliminating missing values, normalizing the cases and other text pre-processing activities. Later, with the help of Tf-idf Vectorizer and count vectorizer feature extraction is performed. In the next process system is going to train the classification models on the training set and predict the outcomes on the test set. The classification approach requires a labelled dataset to train a model for the environment it is working on. The unavailability of the labelled dataset is a major limitation in the classification approach. To overcome the problem of the labelled dataset, we propose an unsupervised learning model combining long short-term memory (LSTM) networks and auto encoder (LSTM-auto encoder) to distinguish spam reviews from other real reviews.

4.2 Proposed Method

The system learns the sequences of words and sentences exist in the review. The LSTM autoencoder has been used for this purpose. The reason behind using an LSTM autoencoder is it preserves the long sequences exist in the reviews. The loss between the input and output of LSTM autoencoder is taken as a feature to identify the real and spam review.

For each review Ii (i1, i2, …, it), the actual and predicted output can be represented as Ii (i1, i2, …, it) and Oi (o1, o2, …, ot) respectively. Using the Mean Square Error (MSE), the reconstruction loss (Rloss) for the review Ii can be evaluated as:

$$Rloss = Oi - Ii$$

Once the reconstruction loss is calculated between the input Ii and output Oi of LSTM autoencoder, the loss is compared with a threshold, to fnd whether the review is fake or real.. The reconstruction loss obtained from autoencoder is passed which grouped each review into either a real or a spam review. These predicted labels are then compared with the actual labels of reviews to find the overall accuracy of the model. The basic 1 layer LSTM network is shown below:

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 731-737 www.ijaem.net ISSN: 2395-5252

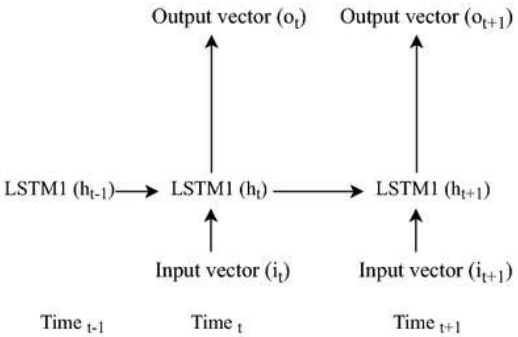


Figure 4.2: 1 Layer LSTM Network

LSTM is a type of the Recurrent Neural Network structure which has recently enhanced the design of RNNs.

LSTM solves vanishing gradient points by substituting the self-linked hidden layers with memory units. The memory units are utilized for storing long-range information of input data when processing. The process of data handling can take place in forward direction. This disregards backward production, so it reduces the performance of the system.

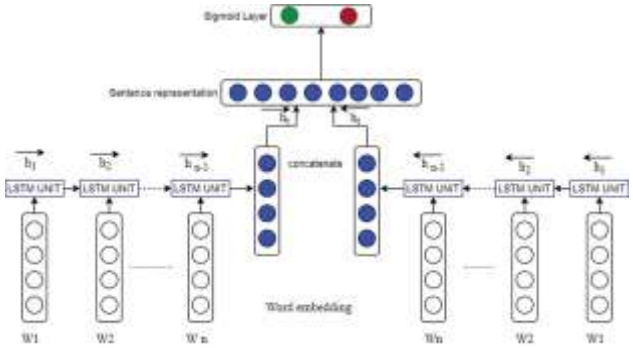


Figure 4.3: Bi-LSTM diagram

4.3 Proposed Algorithm

Steps in proposed systems are as follows:

1. Importing the important libraries and reading the dataset, which contains reviews about different hotels.
2. Dropping the column hotel name, polarity and source.
3. Apply the label encoder to encode the dataset column deceptive.
4. Applying the pre-processing techniques. Following techniques are used as a pre-processing model:
   A) Removing the special characters and symbols.
   B) Remove the digits with hash (#) Symbols. Adding for efficient pre-processing.
   C) Apply some string processing steps.
   D) Apply lambda function for cleaning data.
   E) Apply Count Vectorizer and TF-IDF Vectorizer.
5. Apply Feature Extraction Techniques.
6. Split the dataset into train set and testset.
7. After splitting apply pad sequencing to train data and test data.
8. Apply Glove Embedding for word-word co-occurrence.
9. Create a vector space for embedding size of 100 words.
10. Apply Soft Attention model of LSTM.
11. Train the LSTM model using training dataset. For training the model we select Epoch size= 15 and Batch size =32.
12. After training, the model is evaluated and results are calculated. The accuracy of proposed model is found to be highest.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 731-737 www.ijaem.net ISSN: 2395-5252

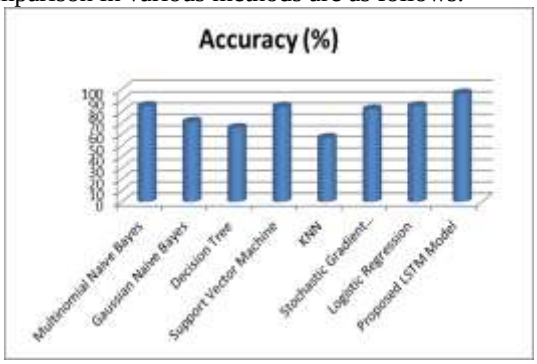
IV. RESULTS AND CONCLUSION

Table below shows the results of proposed and all Base Classifiers in terms of accuracy.

Algorithms	**Accuracy (%)**
Multinomial Naive Bayes	85.25
Gaussian Naive Bayes	71.25
Decision Tree	65.75
Support Vector Machine	84.75
KNN	57.00
Stochastic Gradient Descendent	82.25
Logistic Regression	85.25
Proposed LSTM Model	**96.50**

Table 6.1: Comparisons of Accuracy for Existing and proposed models.

Chart for showing accuracy comparison in various methods are as follows:



In many online sites, there are options for posting reviews, and creating scopes for fake paid reviews or untruthful reviews. These reviews can mislead the general public and put them in a situation to believe the review or not. Machine learning techniques have been introduced to solve the problem of spam review detection.

Much current research has focused on supervised learning methods, which require labeled data - an inadequacy when it comes to online review. Our work in this thesis is to detect any deceptive text reviews. In order to achieve that we have worked with labeled data and proposed deep learning methods for spam review detection which includes Long Short-Term Memory (LSTM).

We have also applied some basic machine learning classifiers such as Nave Bayes (NB), K Nearest Neighbor (KNN) and Support Vector Machine (SVM) to detect spam reviews and finally, we have shown the performance comparison for both traditional and deep learning classifiers.

REFERENCES

[1]. N. Jindal and B. Liu, "Opinion spam and analysis," Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 219–230. ACM, New York, NY, USA (2008).

[2]. J. Fontanarava, G. Pasi and M. Viviani, "Feature Analysis for Fake Review Detection through Supervised Classification," 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Tokyo, 2017, pp. 658-666, doi: 10.1109/DSAA.2017.51.

[3]. Fire M, Goldschmidt R, Elovici Y. Online social networks: threats and solutions. IEEE Commun Surv Tut 2014; 16(4):2019–36. https://doi.org/10.1109/ COMST.2014.2321628.

[4]. Krombholz K, Hobel H, Huber M, Weippl E. Advanced social engineering attacks. J Inf. Secur Appl 2015; 22:113–22. https://doi.org/10.1016/j.jisa.2014.09.005.

[5]. Conti M, Poovendran R, and Secchiero M. Facebook: detecting fake profiles in on-line social networks. In: Proceedings of the international conference on advances in social networks analysis and mining, 2012 August 26. IEEE; 2012. p. 1071–8. https://doi.org/10.1109/ASONAM.2012.185.

[6]. Fire M, Katz G, Elovici Y. Stranger's intrusion detection-detecting spammers and

fake profiles in social networks based on topology anomalies. Hum J 2012; 1(1):26–39.

[7]. Yu H, Kaminsky M, Gibbons PB, Flaxman AD. SybilGuard: defending against Sybil attacks via social networks. IEEE/ACM Trans Networking 2008; 16(3):576–89. https://doi.org/10.1109/TNET.2008.923723.

[8]. Yu H, Gibbons PB, Kaminsky M, and Xiao F. SybilLimit: a near-optimal social network defense against Sybil attacks. In: IEEE symposium in security and privacy, 2008 May 18. IEEE; 2008. p. 3–17. https://doi.org/10.1109/SP.2008.13.

[9]. Danezis G, Mittal P. Sybil infer: detecting Sybil nodes using social networks. In: Proceedings of NDSS, February. The Internet Society; 2009.

[10]. Tran DN, Min B, Li J, Subramanian L. Sybil-resilient online content voting. In: Proceedings of the 6th USENIX symposium on networked systems design and implementation. 2009 April 22, 9. Berkeley, CA, USA: Association; 2009. p. 15–28.

[11]. N. Jindal and B. Liu, "Opinion spam and analysis," Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 219–230. ACM, New York, NY, USA (2008).

[12]. M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT '11), vol. 1, pp. 309–319, Association for Computational Linguistics, Portland, Ore, USA, June 2011.

[13]. J. Fontanarava, G. Pasi and M. Viviani, "Feature Analysis for Fake Review Detection through Supervised Classification," 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Tokyo, 2017, pp. 658-666, doi: 10.1109/DSAA.2017.51.

[14]. S. Shojaee, M. A. A. Murad, A. B. Azman, N. M. Sharef and S. Nadali, "Detecting deceptive reviews using lexical and syntactic features", Proc. 13th Int. Conf. Intelligent Syst. Design Appl., pp. 53-58, Dec. 2013.

[15]. S. P. Rajamohana, K. Umamaheswari and S. V. Keerthana, "An effective hybrid cuckoo search with harmony search for review spam detection", Proc. 3rd Int. Conf. Adv. Electr. Electron. Inf. Commun. Bio-Inform. (AEEICB), pp. 524-527, Feb. 2017.

[16]. C. Catal and S. Guldan, "Product review management software based on multiple classifiers", IET Softw., vol. 11, no. 3, pp. 89-92, Jun. 2017.

[17]. M. Ott, C. Cardie and J. T. Hancock, "Negative deceptive opinion spam", Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol., pp. 497-501, 2013.

[18]. J. K. Rout, S. Singh, S. K. Jena and S. Bakshi, "Deceptive review detection using labeled and unlabeled data", Multimedia Tools Appl., vol. 76, no. 3, pp. 3187-3211, Feb. 2017.

[19]. X. He, X. GAO, Y. Zhang, Z.-H. Zhou, Z.-Y. Liu, B. Fu, et al., "Intelligence science and big data engineering. Big data and machine learning techniques", Proc. 5th Int. Conf. (IScIDE), vol. 9243, pp. 29-42, Jun. 2015.

[20]. K. Goswami, Y. Park and C. Song, "Impact of reviewer social interaction on online consumer review fraud detection", J. Big Data, vol. 4, no. 1, pp. 1-19, Dec. 2017.

[21]. G. S. Brar and A. Sharma, "Sentiment analysis of movie review using supervised machine learning techniques", Int. J. Appl. Eng. Res., vol. 13, no. 16, pp. 12788-12791, 2018.

[22]. Hotel Reviews Dataset, Jun. 2019, [online] Available: https://www.yelp.com/dataset.

[23]. E. I. Elmurngi and A. Gherbi, "Unfair reviews detection on Kaggle reviews using sentiment analysis with supervised learning techniques", J. Comput. Sci., vol. 14, no. 5, pp. 714-726, May 2018.

[24]. A.Mukherjee, B. Liu and N. Glance, "Spotting fake reviewer groups in consumer reviews", Proc. 21st Int. Conf. World Wide Web (WWW), pp. 191-200, 2012.

[25]. P. Kaghazgaran, J. Caverlee and M. Alfifi, "Behavioral analysis of review fraud: Linking malicious crowdsourcing to Kaggle and beyond", Proc. Int. AAAI Conf. Web Social Media, vol. 11, 2017.