2020 23rd International Conference on Computer and Information
Technology (ICCIT), 19-21 December, 2020

A Supervised Machine Learning Approach to Detect Fake Online Reviews

1st Rakibul Hassan
Dept. of computer science & engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
rakibul.hassan@ece.ruet.ac.bd

2nd Md. Rabiul Islam
Dept. of computer science & engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
rabiul.cse@gmail.com

***Abstract*—With increasing use of internet, online business platforms are becoming the largest market place of the world. Purchase of online product is heavily dependent on user reviews. Some dishonest groups of people misuse this fact by posting fake reviews to promote their own products or demote their competitors. Detection of fake online reviews can be considered as a binary classification task that models a classifier to tell whether a review is fake or true. In this paper, we have developed an effective supervised machine learning approach to classify fake online reviews using a dataset that contains hotel reviews from online websites.**

***Index Terms*—Online reviews, supervised learning, support vector machine, naive Bayes, logistic regression, Empath, TF-IDF, sentiment polarity.**

I. Introduction

Purchasing online products is one of our daily activities. Now-a-days, we can get almost everything from various online market places. When we think about purchasing something, almost everyone of us first check the product in websites like Amazon, AliExpress, eBay etc. In case of traveling: hotel booking, purchasing air tickets and all forms of other tasks also can be done with the help of online service providers. But, as we can't know physically, what products or services we are purchasing, we check what other people tell about the services or products. Thus, online reviews play a very crucial role in decision making while purchasing products online. This also creates opportunity for some groups of bad people to deceive people with fake comments and reviews. They can post fake reviews for the promotion of their goods or to demote the products of the competitor. That is why, detecting fake online reviews is very important for both the users to get benefited from reviews as well as the companies to maintain their goodwill to the consumers.

Detecting fake online reviews is basically a binary classification problem. Many researchers are working on it to auto detect the fake reviews using various machine learning techniques. Most of the researches are based on supervised learning. Some semi-supervised and clustering approaches are also taken by some of the researchers. All of these studies, mainly focus on two aspects: reviewers and reviews. Detection system based on only reviews is call content-based study. Classification tasks are done using the features extracted from the content of the reviews that includes what is written in the text. The reviewer-based study is also called user-behavior based classification. How often a reviewer reviews, how varieties of products are reviewed by one reviewer and also the timing and group connections of the reviewers are analyzed here. One of the fundamental problems of all these studies is data labeling. It is very difficult to label the dataset for human by watching the content. This can create a garbage out of garbage situation if reliable labeling is not available. In such cases, semi-supervised or clustering approaches are suggested by many researchers. But, with reliable labeling, supervised classification approaches always perform better.

In this research paper, we have introduced some supervised machine learning classification techniques to detect fake online reviews with a good accuracy. Classifications are done using content-based features. We have used term-frequency and inverse document-frequency (TF-IDF), Empath categories (similar to linguistic word count LIWC) and sentiment polarity as our features. For classification, we have used several sets of classifiers like logistic regression classifier, Naive Bayes classifier and support vector machine (SVM).

In the following section II, related works are discussed. Our proposed approaches are described in Section III. Results and findings of our research are shown in Section IV. Conclusion and future work are discussed in section V.

II. Related Works

Many approaches have been taken by the researchers to detect fake online reviews. Those approaches have brought significant improvement to separate fake and truthful reviews effectively. Sun et al. [1] have categorized the approaches into two aspects.

- Content based study.
- User-behavior based study.

Content based study focuses on the context of the reviews and user behavior or characteristics based study focuses on the individual's characteristics who is posting the review. Many approaches have been taken based on content-based study. Jindal et al. [2] divided fake reviews into two types.

- Deceptive reviews: Posting fake positive reviews deliberately to promote the selection of a certain product. Posting fake negative reviews to hamper the reputation of certain goods and putting negative impact on the sales.

978-0-7381-2333-2/20/$31.00 ©2020 IEEE

- Destructive reviews: This kind of reviews are unrelated advertisements which are not related to the review motive.

Destructive reviews aren't that much harmful. People can understand from reading that they are not relevant to the product. Hence it doesn't impact the decision making. But deceptive ones play huge role in decision making and detecting deceptive reviews is main task of content-based study.

Ott et al. [3] proposed three different type of features to perform classification. These types are- identification of genre, detecting psycholinguistic behavior and categorization of text.

1) Identification of genre: The parts-of-speech (POS) distribution of text was analyzed by Ott et al. [3]. They suggested that, POS tags can be used for feature representation of classification.
2) Detecting psycho linguistic behavior: The psycho linguistic method considers psycho linguistic meanings as important features of a review. Linguistic Inquiry and Word Count (LIWC) tool was introduced and used by Pennebaker et al. [5] to create their features for the reviews.
3) Categorization of text: Ott et al. [3] used n-gram feature in fake review detection. They showed that, n-gram feature has good importance in deceptive review classification.

Ott. et al. [3] generated a gold standard dataset using Amazon Mechanical Turk (AMT). They achieved around 89% accuracy using only the positive fake reviews. They used parts-of-speech and n-gram features. As the dataset was small and there were no deceptive negative reviews the accuracy didn't described the whole picture. Lately, Ott. et al [4] updated the dataset with deceptive negative reviews and accuracy of classification decreased from the previous performances.

Feng et al. [6] used unlexicalized syntactic and lexicalized features by constructing sentence parse tree to detect fake reviews. They experimentally showed that accuracy of prediction can be improved by using deep syntactic features.

Li et al. [7] experimented a variety of generic signals that can make impact on the fake review detection. They also found that combination of general features such as LIWC or Parts-of-speech with word frequency count is more useful than word frequency count alone as features. Metadata about reviews such as date, time, rating, length of review are also used as important features by some of the researchers.

User-behavior based analysis is also done by many researchers. Lim et. al [8] tried to detect product review spammer using rating-providence based characteristics of the reviewers. They used amazon product reviews with user and product information for fake reviewer detection. They have discussed about general deviation of rating of the spammers and product group-based studies. They have discussed the following fake review rating behaviors.

a) Providing bad rating too often: Professional fake review writers usually post more deceptive reviews than genuine ones. For example, we can think about a specific product that has average rating of 4.5 in a scale of 5. But it appeared that one reviewer has provided 0.0 rating. Exploring some of the other reviews of that reviewer if we find out that he most often provides unfair rating like this, then, he may be considered as a review spammer.

b) Providing very good rating to the products of native region: It happens sometimes that, user posts fake reviews to promote products of native region. This kind of review manipulation is more often seen in movie reviews. Suppose, in a popular movie website like IMDB, we have a Bollywood movie that has a rating 5 out of 5. But, we discovered that most of the reviewers are South Asian. We can detect this type of spamming using reviewer's ip address.

c) Providing reviews on varieties types of products: Every individual has product specific attraction. One individual is generally not attracted to all kinds of products. If someone appears to provide reviews in varieties of products, it is a deviation from the general behavior. We can intuit him as a spammer.

Mukherjee et al. [9] discussed about various aspect of review-based and reviewer-based studies on real world yelp dataset. They suggested that, only analyzing reviews will not provide good performance on real world data. They proposed to use various user behavior-based features like maximum number of reviews, reviewer deviation, maximum content similarity (MCS), percentage of positive reviews etc. for better classification.

Detecting fake online reviews can be considered as a binary classification problem and most popularly supervised approaches are adopted for this task [6]. These supervised approaches performs very well if we train on large dataset of labeled instances. The dataset should have enough examples from both positive and negative classes [10]. For labeling the dataset ground truth is measured from human observation, rating behaviors, seed words, helpfulness vote etc. Sun et al. [1] implemented a model that produces classification output through bagging approach. The bagging model includes three classifiers which are $BIGRAMS_{SVM}$ classifier, product word composition classifier (PWCC) and $TRIGRAMS_{SVM}$ classifier. To predict the polarity of the review, they used a product word composition classifier. Using Convolutional Neural Network (CNN) for product word composition and bagging with two SVM classifier they achieved F-score value 0.77.

When reliable labeling is not available, semi-supervised models can perform well. Rout. et al. [10] proposed several semi-supervised classification techniques on Ott. et al [4] dataset that contains both positive and negative deceptive examples. They used semi-supervised approaches with Expectation maximization, Co-training, Positive Unlabeled Learning and Label Propagation and Spreading algorithms for training. They performed classification with several classifiers including Logistic Regression, Random Forest, K-Nearest Neighbor and Stochastic Gradient Descent. They achieved highest of 84% accuracy using these semi-supervised machine learning approaches.

Hassan et. al [11] implemented both semi-supervised and

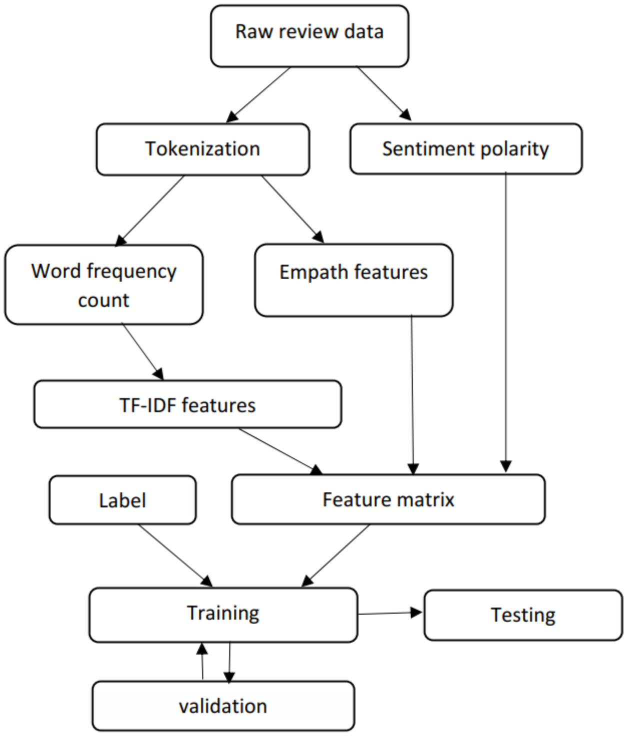


Fig. 1: Proposed classification model

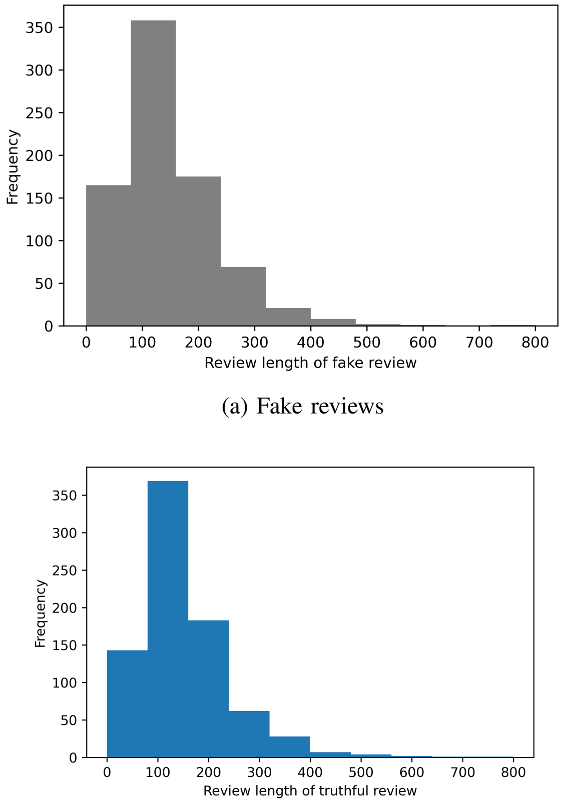
supervised machine learning techniques to detect fake online reviews on the same data set used by Rout. et al [10]. They have shown that, on that dataset supervised classifiers perform better. Using supervised Naive Bayes classifier, they got an accuracy of 86%. They concluded that the dataset can be considered more reliable for supervised classification.

III. Proposed Work

A. Dataset description

In this research, a gold standard hotel review dataset developed by Ott. et.al [4] is used. This dataset is also used by Rout et al. [10] and Hassan et al. [11]. Firstly, Ott et. al [3] developed a dataset which only contained positive fake reviews. The fake review data was collected using Amazon Mechanical Turk (AMT). The targeted hotels were the most popular 20 hotels in Chicago area of United States, listed on TripAdvisor website. Through the AMT platform, the marketing department managers of those hotels requested users to write positive reviews of their hotel. Thus, 400 fake reviews of positive sentiment were collected. 400 truthful reviews were taken from several websites like TripAdvisor and Yelp. The classification result was good but dataset was imbalanced as deceptive negative reviews were not collected. Lately, Ott.et al. [4] improved the existing dataset by collecting more 400 deceptive reviews with negative sentiment. Also, more 400 truthful reviews were added to balance the dataset.

Finally, we have 800 reviews which are deceptive and 800 reviews that are truthful. For evaluation, a tag of '1' indicates truthful reviews, whereas '0' indicates the deceptive ones. In the dataset, from 800 truthful reviews half are written with a negative sentimental polarity and rest half includes positive sentimental polarity. Similarly form deceptive ones, 400 are positive and rest 400 have negative sentiment polarity.



(a) Fake reviews

(b) Truthful reviews

Fig. 2: Review length distribution

For evaluation, the dataset is splitted in a constant partition. From the 1600 examples in the dataset, three sets of examples are separated: the training set, validation set and the test set. The test set contains 20% of total examples. The train-validation set contains the rest 80% examples. The train-validation set is splitted with a ratio 75:25 to obtain the train set and validation set. Random sampling is used for partitioning each set.

B. Proposed methodology

To develop a model that can classify deceptive and truthful reviews effectively we have used TF-IDF, Empath and sentiment polarity as our features. Hassan et. al [11] used word frequency count as their features and trimmed the most frequent and less frequent ones to develop a dictionary. They suggested that, as articles, prepositions etc. are top frequent words and they don't carry that much information those can be excluded. Another way of using all of the words but giving less importance to those kinds of less meaningful words is using term frequency and inverse document frequency (TF-IDF). Li et al. [7] and some other researchers used TF-

IDF and found that TF-IDF works well to classify fake online reviews. Rout et al. [10] used linguistic word count (LIWC) as one of their features. We have used similar feature called Empath that can be generated using open source tool developed by Fast et. al [12]. Fast et al. [12] showed that Empath features are highly correlated (r=0.906) with LIWC. Moreover, it is data driven that uses deep learning for category development. Empath produces a large set of categories, and it can produce and validate new set of categories if we need using unsupervised language modeling. Also, the effectiveness of empath categories was analyzed on dataset developed by Ott. et al. [4], that we have used in this research. Another feature of our model is sentiment polarity that tells about whether a review is given positively or negatively. This feature is used by both Hassan et al. [11] and Rout et al [10]. We have analyzed the review length feature used by Hassan et al. [11] and found that, the distribution of this feature in our dataset is similar for both fake and truthful reviews. The average length is 146 for fake reviews and 150 for truthful reviews. Hence we haven't used this feature as it may put very light impact in the performance. Review length distribution for fake and truthful reviews are shown in Fig. 2. In case of other real world datasets, it may be left skewed for fake reviews as professional spammers typically try to post short reviews.

Fig.1 shows our proposed model to detect fake online reviews. Our classification process uses the following steps.

1) We need to take all review data to from the information directory. Then, we retrieve sentiment polarity as well as the labeling of each review. We tokenize the words of the review and generate the dictionary.
2) From the tokens we count the frequency of each word in a review and calculate the TD-IDF. Meanwhile we use Empath tool for retrieving Empath features. Empath tool gives us around 200 categories.
3) We combine TF-IDF, Empath categories and sentiment polarity to compose the feature matrix. And we shuffle the data randomly before separating train-validation and test set.
4) The feature matrix as well labels of the features are feed to a classifier for training. The validation set is used to remove overfitting and fine-tune different parameters of the classifier. The model created from training is used to analyze the performance with test data set.

Following the model we have described above, 9571 features are extracted as TF-IDF features. Empath extracted categories are 194. With sentiment polarity as another feature we have used in total 9766 features in this proposed system. Here, we have used supervised learning to train the classifier. Rout et al. [10] used semi-supervised learning for training and Hassan et al. [11] used both supervised and semi-supervised learning to train the dataset. We have used logistic regression, Support Vector machine (SVM) and Naive Bayes classifiers as our classifiers.

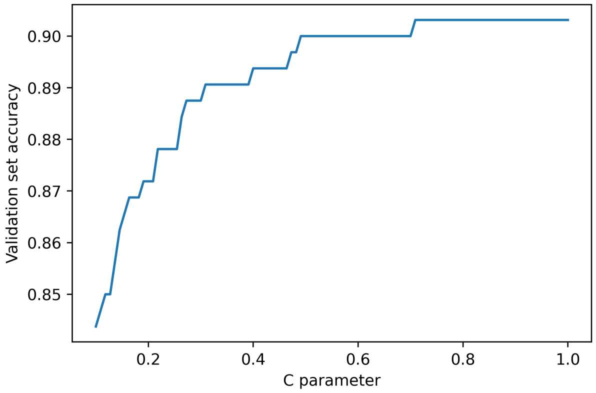


Fig. 3: C parameter tuning for logistic regression

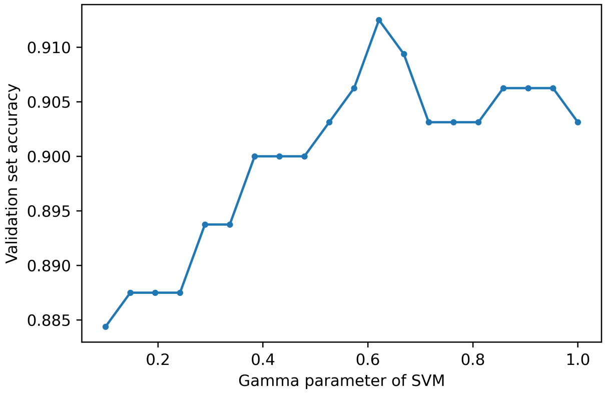


Fig. 4: SVM classifier's gamma parameter tuning

IV. PERFORMANCE ANALYSIS

A. Evaluation Matrices

To evaluate the performance of our proposed techniques and compare with previous work, we have taken accuracy, precision, recall and F1 score as evaluation matrices. As our data set is totally balanced, accuracy is a sufficient metric to compare the performances. F1 score measurement is used to know how effectively our classifier can classify both positive and negative classes. The receiver operating characteristic (ROC) curve is also drawn and area under the curve is calculated to measure the goodness of each classifiers.

B. Results

We have used three classifiers Naive Bayes, logistic regression and support vector machine for training the model. In each case we have used train, validation and test set at a ratio 60:20:20.

Using multinomial Naive Bayes classifier, we have obtained 85.62% validation set accuracy and 84.37% test accuracy. Among 320 classified test examples, we found: 149 true negative, 43 false negative, 121 true positive and 7 false positive classes. The precision here was 0.94 and recall was 0.74. The F1 score we had obtained was 0.83.

TABLE I: Comparative summary of findings

Works	Features	Algorithm	Classifier	Accuracy	Precision	Recall	F1 Score
Rout et al. [10]	Bigrams, Sentiment score, POS, LIWC	Semi-Supervised	Logistic Regression (PU)	0.8375	0.8313	0.8418	0.8365
			K-NN with (EM)	0.8313	0.8063	0.8487	0.8269
Hassan et al. [11]	Word frequency count, review length, sentiment score	Semi-Supervised	Naive Bayes	0.8521	_	_	_
			SVM	0.8134	_	_	_
		Supervised	Naive Bayes	0.8632	_	_	_
			SVM	0.8228	_	_	_
Proposed Work	TF-IDF, Empath categories, sentiment score	Supervised	Naive Bayes	0.8437	0.9453	0.7378	0.8287
			Logistic Regression	0.8843	0.8802	0.8963	0.8882
			SVM	0.8875	0.9050	0.8719	0.8881

With logistic regression classifier, at first, we have fine tuned the C parameter of the classifier using "lbfgs" solver and setting penalty to l2. The validation set accuracy with different C parameters is shown in Fig. 3. Thus we used 1.0 as value of C. With this setup we have achieved 90.31% validation set accuracy and 88.43% test accuracy. Among 320 classified test examples: 136 were true negative, 17 false negative, 147 true positive and 20 false positive. The precision here was 0.88 and recall was 0.90. The F1 score we found here was 0.89.

For SVM classifier we have fine tuned the gamma parameter with keeping 1.0 value for C parameter. We have obtained highest 91.25% validation set accuracy setting gamma to 0.62105. Fig. 4 shows the validation set accuracy with different gamma parameter values. With this set up of C and gamma parameter we have obtained 88.75% test accuracy. This is the highest among all three classifiers we have used. With SVM classifier we found 141 true negative, 15 false positive, 21 false negative and 143 true positive classes. The precision here was 0.91 and recall was 0.87. The F1 score here we found was 0.88. A histogram showing these results is given in. Fig. 5.

C. Result analysis

Hassan et. al [11] got highest 86.32% accuracy with supervised learning using Naive Bayes classifier. They used 80:20 data split ratio for training and testing. Rout et al. [10] achieved 84% accuracy using semi-supervised Positively Unlabeled (PU) learning with logistic regression classifier. Here, with same train-test ratio, our model with Naive Bayes, Logistic Regression and SVM classifiers provides 84.37%, 88.43% and 88.75% classification accuracy respectively. The results clearly show that, our proposed supervised classification approaches can classify fake online reviews effectively with better accuracy on the hotel review dataset. Also the

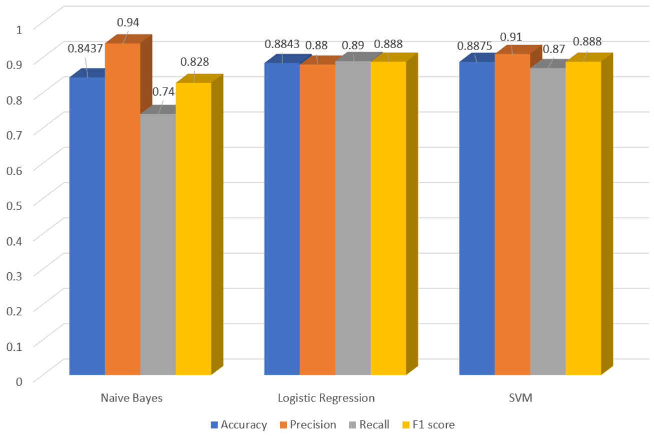


Fig. 5: Histogram showing performances of implemented classifiers

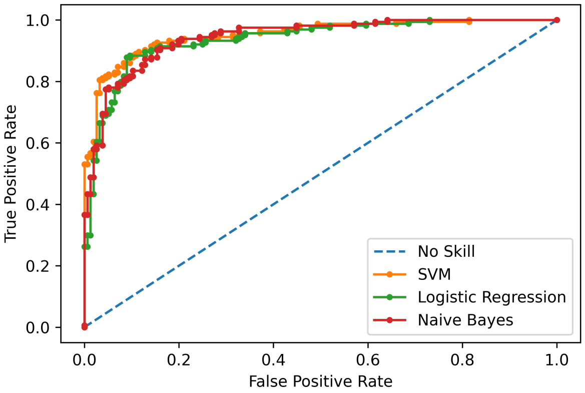


Fig. 6: ROC curve for implemented classifiers

precision, recall and F1 score indicates a good quality of classification. The Receiver Operating Characteristics (ROC) curve of the classification approaches are given in Fig. 6. The area under the curve (AUC) is also large for each classifier. For SVM Classifier it is 0.952, for Logistic regression Classifier the AUC is 0.936 and for Naive Bayes Classifier it is 0.944. Our findings are summarized in Table. I.

V. CONCLUSION AND FUTURE WORK

We have shown some supervised machine learning classification techniques for detecting fake online reviews in this research. We have merged features from some other research works for development of a feature set that can perform better classification. Thus, we have increased the accuracy from previous semi-supervised techniques done by Rout et al. [10] as well as supervised and semi-supervised techniques done by Hassan et al. [11]. We have developed a supervised SVM classifier based classification system that gives 88.75% accuracy. Here, we have worked with just online reviews. In future, user behavior based characteristics can be used alongside review text to develop a model with better classification results. Efficiency and effectiveness of the proposed model can be tested with dataset of larger size. We have done our research only on reviews of English language. It can be implemented on languages like Bangla and others.

REFERENCES

[1] Chengai Sun, Qiaolin Du and Gang Tian, “Exploiting Product Related Review Features for Fake Review Detection,” Mathematical Problems in Engineering, 2016.

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

[3] 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.

[4] M. Ott, C. Cardie, and J. T. Hancock, “Negative deceptive opinion spam,” in Proc. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol., 2013, pp. 497–501.

[5] J. W. Pennebaker, M. E. Francis, and R. J. Booth, Linguistic Inquiry and Word Count: Liwc 2001, vol. 71, Lawrence Erlbaum Associates, Mahwah, NJ, USA, 2001.

[6] S. Feng, R. Banerjee, and Y. Choi, “Syntactic stylometry for deception detection,” in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers— Volume 2, pp. 171–175, Association for Computational Linguistics, 2012.

[7] J. Li, M. Ott, C. Cardie, and E. Hovy, “Towards a general rule for identifying deceptive opinion spam,” in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL ’14), pp. 1566–1576, June 2014.

[8] E. P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, “Detecting product review spammers using rating behaviors,” in Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM), 2010.

[9] A. Mukherjee, “Detecting deceptive opinion spam using linguistics, behavioral and statistical modeling,” in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, Technical Report, Beijing, China, July 2015.

[10] J. K. Rout, A. Dalmia, and K.-K. R. Choo, “Revisiting semi-supervised learning for online deceptive review detection,” IEEE Access, Vol. 5, pp. 1319–1327, 2017.

[11] R. Hassan and M. R. Islam, ”Detection of fake online reviews using semi-supervised and supervised learning,” 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox’sBazar, Bangladesh, 2019, pp. 1-5, doi: 10.1109/ECACE.2019.8679186.

[12] Ethan Fast, Binbin Chen and Michael Bernstein, . ”Empath: Understanding Topic Signals in Large-Scale Text,” May 2016, San Jose, CA, USA 2016 ACM, doi: 10.1145/2858036.2858535.

.