Fake Reviews Detection Using Machine Learning

Swijel Dmello[1], Ridhi Bauskar[2], Apoorva Shet[3]

Information Technology Department, Fr. Conceicao Rodrigues College of Engineering, University of Mumbai, India
[1]swijeld[at]gmail.com
[2]ridhi01bauskar[at]gmail.com
[3]apoorvashet11[at]gmail.com

Abstract: ***Fake Review Detection is a very crucial task in the field of e-commerce and business. It plays a pivoting role in the decision-making and quality assessment of a product. However, this task of reviewing is carried out manually by humans or by a random review checker mechanism. In order to review false positive and false negative reviews, a fake review dataset has been used. Due to the advancements in the field of Machine Learning and Natural Language Processing, these algorithms could be leveraged to detect fake reviews with high accuracy and in a short amount of time which would save a huge amount of manual effort. This paper proposes an approach to detecting fake reviews through the various advanced machine learning techniques like Support Vector Machines, Decision Trees, etc. The system put forward is a web-based solution that provides an accurate result whether the given review is valid or not.***

Keywords: Fake Review, Sentiment Analysis, Machine learning, SVM, KNN, GBB, NLTK, SGD, Decision Tree, Logistic Regression, MNB

1. Introduction

The Social Web has made an intense change in the existence of everybody these days through communicating their perspectives on the web. The size of the client-created content is developing quickly. E-Commerce gives a polished experience to web-based clients. The sites permit their clients to give their criticism as surveys on their destinations. Over 90% of shoppers before buying any item or utilizing any assistance, it has turned into a propensity for perusing web surveys for a dynamic reason. Around 40% to 70% of audits given in Online locales are found as false surveys. The new clients as well as the current clients think about the surveys as their significant wellspring of data. Every one of the reviews composed is false. A portion of the audits are spam as they are composed for certain advantages like an advertisement for their item, promotion of their item or administration, essentially spreading information or now and then out of these phoney surveys might even get Financial Profits. Hardly any business ventures utilize an individual of letters to draw up manufacturing positive surveys on their items and corrupt negative audits on their adversary's items. Thus, simply accepting these internet-based surveys and settling on choices might turn out badly because not all reviews are authentic. It deludes clients and the absence of control in spam data spreading. Dimensionality decrease should be done to build the outcome. The Review Dataset is profoundly unique and veracity in nature as it is brimming with text information.

The purpose of this research is to apply a machine-learning algorithm to predict whether reviews are genuine or not using a dataset of hotel reviews. Using the Support Vector Classifier algorithm, which is one of the ensemble learning approaches' properties. We devote our efforts to developing a system that can detect phoney reviews on websites where they are most likely to appear, as well as in real-time data such as social media apps. In comparison to other algorithms and approaches such as Decision Tree, Logistic Regression, KNN, and others, the SVC model has the best combination of prediction performance and processing time.

2. Related Work

In a study based on Detecting Fake Reviews Using Machine Learning Techniques algorithms like Sentiment Analysis, NB, DT-J48, KNN-IBK, and SVM were implemented on the movie review dataset here collusion and manipulation were discussed. In [7] The Sixth International Conference on Data Analytics, here the paper classifies reviews as negative or positive polarity using 5 Machine Learning algorithms, Support Vector Machine (SVM), KStar, K-Nearest Neighbors, Decision Tree and Naïve Bayes for sentiment classification of reviews using two different movie review datasets.

In [10] Deceptive review detection using labelled and unlabeled data (Springer). Review text-based, as well as reviewer-based methods, have been used which are automated methods used to detect spam reviews. Best feature sets are acquired to reject reviews of spam or nonspam. Two different datasets have been used and unsupervised and supervised learning techniques are used to review both the data sets separately. Finally, a comparison between different feature set analyses based on review text, sentiment scores, reviewer feature, and combined method.

The integration of a bag of words, as well as words of context and consumer emotions, is employed in [10] focuses on refining the fake review detection approach utilizing two neural network models. (1) n-grams, (2) word embeddings, and (3) multiple lexicon-based mood indicators are among the set's features. The performance is unaffected by the dataset's sentiment polarity or product category.It performs admirably across all datasets.

The requisites to detect false reviews can be discussed in the Journal of Xi'an University of Architecture and Technology 2020 Groceries and Home Appliances Reviews Using SVM. In this [8] paper, the system's accuracy is lower. Only a few platforms, not all, can be used to detect fake reviews. Only labelled datasets can be used, not unlabeled datasets. Fake reviews can only be detected using a few techniques[24].

3. Dataset Description

This paper utilizes the root dataset of hotel reviews against different reviews classification models mentioned in this paper. This paper utilizes the publicly available dataset provided by Cornell University which is usually considered the gold standard data for research done in the domain of Text Analysis. This dataset is known as the “Deceptive opinion spam corpus”. The characteristics of the dataset are deceptive, hotel names, polarity, source and the text. The dataset is in CSV format and consists of 1600 records.

This corpus contains

- 400 truthful positive reviews from TripAdvisor [1].
- 400 deceptive positive reviews from Mechanical Turk [1].
- 400 truthful negative reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor and Yelp [2].
- 400 deceptive negative reviews from Mechanical Turk [2].

Each of the datasets consists of 20 reviews for each of the 20 most popular Chicago hotels.

4. Methodology

a) Data Pre-Processing

The first step in the data preprocessing stage is to mould the present data into a relevant form by removing stop words, and punctuations apart from this we also convert all the text into lower case, the steps are necessary before using the n-gram model. This data preprocessing helps to exclude the information that is outliers and may add noise in the model training process and aids the model to focus on the relevant data points. After we have removed the punctuations, and stop words and also converted the words into lowercase, the current approach applies lemmatization in order to extract the keywords from the raw data for the next steps in the process.

b) Feature Extraction

The most pivoting factor to achieve better accuracy on classifier models is to provide relevant features to them. Hence, the feature extraction stage plays a very crucial role. The efficiency and performance of a Machine Learning model are drastically affected by irrelevant data or outliers. Hence, eliminating these types of data points becomes necessary before we extract the key features so that the model receives effective and relevant data. The characteristic of the TF-IDF vectorizer is to generate a bag of words from each review and these groups of words are further used for tokenization.

c) Classifier model construction and testing

A dataset that is comparatively small is easy to handle for training and so it's advised to use smaller datasets for training. Reviews used for testing purposes already have a label of being real or fake. The classification model plays the role of just reviewing the dataset used. Another dataset which is a tester dataset is used for the classification model after the training process of the classifier model. The different machine learning algorithms which we used for model construction are, Decision tree algorithm, support vector machine, Gaussian naive Bayes, k-nearest neighbour, Multinomial Naïve Bayes and logistic regression.

d) Model Evaluation and Training

Checking if the model is ready for deployment is a crucial part of the process. For all of the models we used for comparative analysis, we used Grid Search to determine the optimal parameters.

For data handling and preprocessing, the following tools were used:

1) NumPy, pandas, pickle, and matplotlib [15].
2) Visualization base map.
3) Sklearn for model evaluation and classification models [14].
4) For model deployment, I used Django with HTML, CSS, and the web [16].

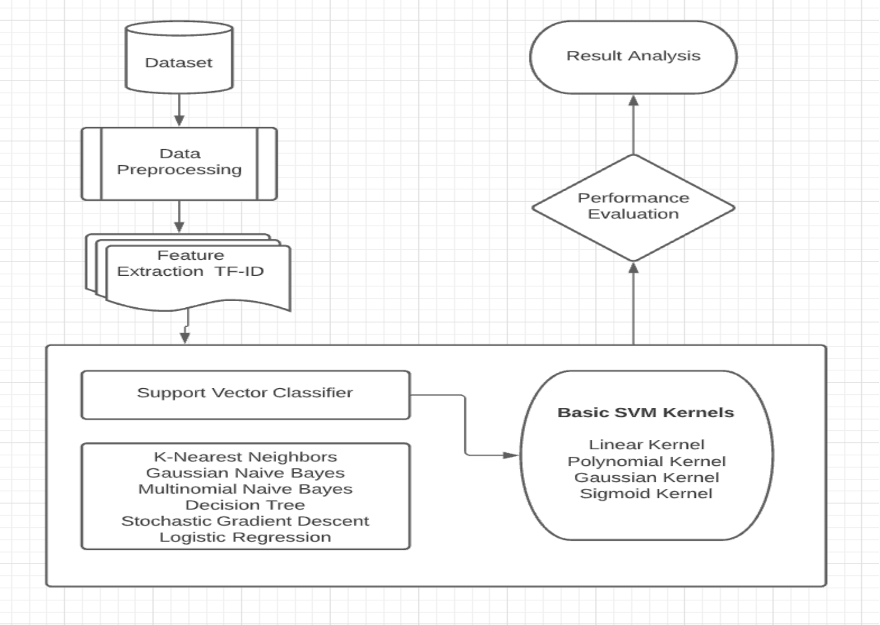

Figure 2.1: System Architecture

Fig. 2.1 explains the proposed model pre-processing and selection workflow this paper has followed and also concentrates on the various SVM kernels used.

5. Implemented Algorithms

a) Decision Tree (D-Tree):

By using non-parametric supervised learning methodologies, the Decision Tree algorithm produces prediction models [14]. Here, from the training data, simple decision rules are presumed for the creation of training models that can predict the classes and the values of actual variables. The accuracy resulting from using the Decision tree is 70.5%.

b) B. K-Nearest Neighbor (KNN):

By measuring the Euclidean distance between two points, the KNN algorithm captures similarities between them. For this study, a simple KNN class is called with a k value assigned to the number of notable crimes in the dataset, and the results are estimated accordingly. The accuracy resulting from using KNN is 74.75%.

c) Naïve Bayes (NB) Algorithms:

Naive Bayes is a multi-class binary classification method. Hence this also suits the given problem statement. In this paper, we cover two types of Naive Bayes Algorithms mentioned below:

- **Gaussian – Naive Bayes:**
  Gaussian Naive Bayes is a variation of Naive Bayes that handles continuous data and follows the Gaussian normal distribution. The Gaussian or Normal distribution is used since you only need to estimate the mean and standard deviation from your training data. The accuracy resulting from using this algorithm is 65.75%.

- **Multinomial Naive Bayes:**
  The Multinomial Naive Bayes classifier generally performs well for classification problems that contain discrete features. The multinomial distribution in this algorithm usually takes in consideration integer feature counts. The Naive Bayes Algorithm firstly finds the tag of a text and then calculates the probability of each and every tag for a given text and then outputs the highest tag. The accuracy resulting from using this algorithm is 85%.

  Stochastic gradient descent is an optimization approach for reducing a model's loss across a training dataset. For each class C, a binary classifier is learned to distinguish it from all other C-1 classes. The accuracy resulting from using this algorithm is 87.25%.

d) Stochastic Gradient Descent (SGD):

Stochastic gradient descent is an optimization approach for reducing a model's loss across a training dataset. For each class C, a binary classifier is learned to distinguish it from all other C-1 classes. The accuracy resulting from using this algorithm is 87.25%.

e) Support vector machine (SVM):

The dataset expected to train an SVM model should be divided and organized into two different categories. Hence, the model is moved for construction after it is entirely trained. Furthermore, the major goal of the SVM model is to identify and map a new data point to a predefined category. Apart from this, the SVM model should also maximize the margins between the two defined categories [13]. The end result of the SVM model is to find a hyperplane that segregates the given data points into two mutually exclusive categories. A hyperplane is a boundary that helps in decision making. The data point falling on different sides can be distributed to different classes. In SVM, a kernel takes the features as input and creates the linearly separable data in a higher dimension. There are multiple kernels that can be used with SVM's as mentioned below, however, the best-suited kernel for a large dataset of review articles would be RBF because of its best efficiency in handling multiple data points. SVM with an RBF kernel yields a maximum accuracy of 89.25% on the dataset used.

1) Gaussian Kernel (RBF):

RBF kernels are the most generalized form of kernelization and are the most widely used kernels. Its features are similar to the Gaussian distribution.

$$K(X_1, X_2) = exp(-\frac{||X_1 - X_2||^2}{2\sigma^2})$$

The RBF kernel is a function of two points X_1 and X_2 that finds the similarity or how close they are to each other. Here, 'σ' is the variance and the hyperparameter. RBF value can go only till 1 this happens when (X1 - X2) equates to 0. This scenario occurs when both the points X1 and X2 are the same which implies that there is no distance between them and so they are similar. When the points are separated by a larger distance, then the kernel value becomes less than 1 and close to 0 which would mean that the points are dissimilar, hence classified as different classes. Based on the threshold, points can be classified in your major classes.

The maximum accuracy resulting from using this kernel with SVM is 89%.

2) Polynomial Kernel:

It is used to learn non-linear models by representing the similarity of vectors (training samples) in a feature space over polynomials of the original variables. The polynomial kernel, on the surface, looks at the supplied features of input samples to identify their similarity, as well as combinations of these features. The polynomial kernel's equation is as follows:

$$K(X_1, X_2) = (a + X_1^T X_2)^b$$

The constant term is 'a', the kernel degree is 'b', and the two data points are X1 and X2. Because greater degrees tend to surpass NLP difficulties, the most typical degree is b = 2 (quadratic).

This kernel has an accuracy of 88%.

3) Linear kernel:

The Linear Kernel is employed when the data points are linearly separable, that is when they can be split by a single line. It's one of the most popular kernels out there. It's commonly used when a Data Set has a disproportionately large number of Features. When there are a lot of features, the linear kernel is ideal. This is because moving the data to a higher-dimensional space has no effect on performance. The number of occurrences (documents) and unique characteristics (words) in text categorization are both enormous. Linear Kernel takes less time to train than other algorithms because just the optimised hyperparameter 'C' is required while other models have various different parameters as well.

The accuracy with this kernel is 88%.

4) Sigmoid Kernel:

The Sigmoid Kernel is derived from the Neural Networks field, where the bipolar sigmoid function is normally used as an activation function for artificial neurons.

$$k(x, y) = \tanh(\alpha x^T y + c)$$

Slope alpha and the intercept constant c are two adjustable parameters in the sigmoid kernel. The value for alpha is 1/N, where N is the data dimension. This function is similar to a two-layer perceptron model that is used in neural networks, this is used as an activation function for artificial neurons in deep learning.

The accuracy with this kernel is 88.5%.

f) Logistic Regression:

The logistic classification model is a binary classification model during which the conditional probability of one of the two possible realizations of the output variable is assumed to be adequate to a linear combination of the input variables that are transformed by the logistic function. Since this paper has only two classes, this algorithm is suitable and may easily be trained thanks to its efficient approaches.

The accuracy using this model was 88.25%.

6. Results and Analysis

Table 3.1: Comparison of different algorithms

S. No.	Algorithm Implemented	Accuracy	Precision	Recall	F1 Score
1	Support Vector Machine	0.8925	0.817	0.8634	0.8894
2	K-Nearest Neighbors	0.7475	0.8561	0.6097	0.7122
3	Gaussian Naive Bayes	0.6375	0.6442	0.6536	0.6489
4	Multinomial Naïve Bayes	0.85	0.914	0.7804	0.8421
5	Decision Tree	0.705	0.7121	0.71219	0.712195
6	Stochastic Gradient Descent	0.8725	0.8877	0.8731	0.8753
7	Logistic Regression	0.8825	0.9072	0.8585	0.8822

Table 3.1 the table above shows the final result obtained using the different models.Confusion matrix of all the models helps in understanding the performance of each model on the given dataset.The results mentioned in the above table are an average of the multiple successive trials done on the dataset. Based on the results in Table 3.1, the graph in Fig. 3.1. The graph has all the algorithms on X axis and accuracy on Y axis. It shows that SVM provides us with the highest accuracy followed by LR, SGD, Naïve Bayes, KNN and finally Decision Tree.

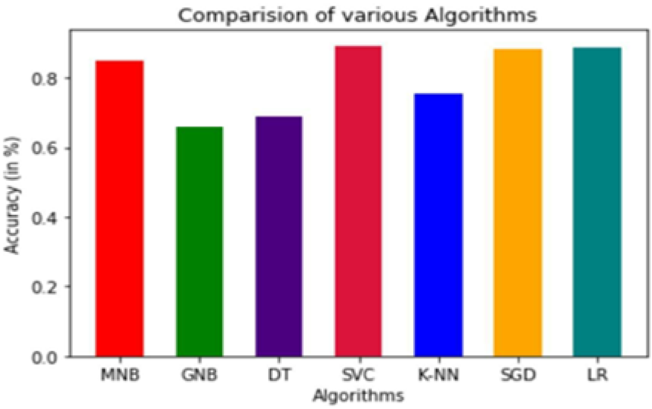

Figure 3.1: Graph of different algorithms

Table 3.2 shows The various parameters like accuracy, precision, recall, F1 Score which we calculated for different kernels.

S. No.	Kernel Implemented	Accuracy	Precision	Recall	F1 Score
1	Gaussian Kernel (rbf)	0.89	0.917	0.8634	0.8894
2	Polynomial Kernel	0.88	0.9067	0.8536	0.8793
3	Linear Kernel	0.88	0.9067	0.8536	0.8793
4	Sigmoid Kernel	0.885	0.9119	0.8585	0.8844

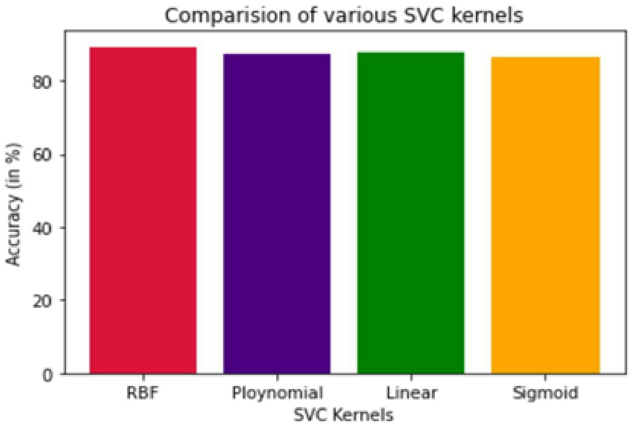

Figure 3.2: Graph of different kernels

On Basis of Table 3.2, Fig. 3.2 it shows the various SVM kernels which we trained and found that RBF kernel achieves the maximum accuracy.Since the data we are analyzing is not symmetric the linear kernel of SVM would not be a perfect fit for the model. Therefore, kernels that take into account unsymmetrical data would perform well compared to the other linear models. Hence, the RBF kernel gives us better accuracy since it creates a non linear hyperplane.

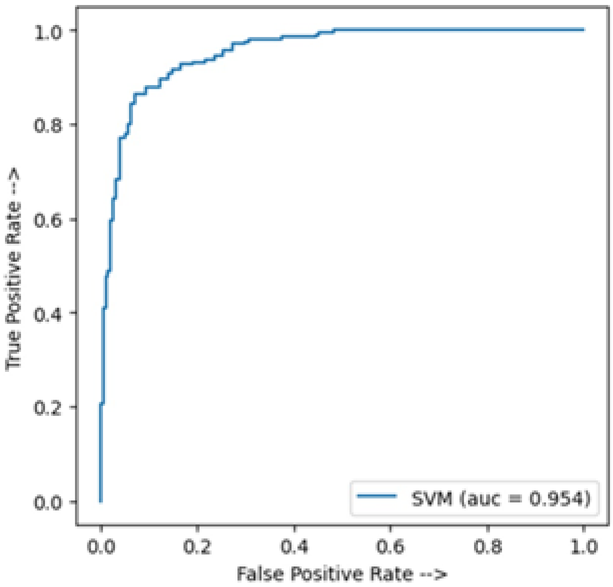

Figure 3.3: Graph of Area under the curve

The above figure shows the area under the curve.

ROC is a probability curve and AUC **is a measure or** degree of separability. It tells the model **how to differentiate between classes. The higher the AUC, the better the model predicts class 0 as 0 and class 1 as 1.**

Here, we got AUC for SVM as 0.945.

7. Conclusions

This paper shows the best Machine Learning model for fake review detection after analyzing all the machine learning models. After the comparison of all kernels of the best model is SVM, all kernels of SVM are tested for comparing their accuracy. The SVM model provides the highest accuracy in the RBF kernel. The highest accuracy achieved is 89%. Fake review detection is an emerging research area that has a scarce number of datasets. There is no data on real-time news or regarding current affairs. The current model is trained using the existing dataset which shows that the model performs well against it.

Utilizing a moderately bigger dataset to prepare the framework can be one of the things to come for our venture. The subsequent stage then, at that point, is to further train and examine the model figure out how the accuracy fluctuates with other datasets and develop it further.

Acknowledgement

We would like to thank our professor and project mentor Mr Unik Lokhande for his constant support and guidance throughout the project. We are very grateful to the Head of the Department of Information Technology Engineering, Mrs Jgruti Save and the college principal, Dr Srija Unnikrishnan for their generosity and co-operation during our project.

References

[1] M. Ott, Y. Choi, C. Cardie, and J.T. Hancock. 2011. 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.

[2] M. Ott, C. Cardie, and J.T. Hancock. 2013. Negative Deceptive Opinion Spam. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

[3] T. J. Ma and D. Atkin, "User-generated content and credibility evaluation of online health information: A meta-analytic study," Telematics and Informatics, 2016.

[4] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance, "What Yelp fake review filter might be doing?" in Proceedings of ICWSM, 2013.

[5] Parikh, S. B., & Atrey, P. K. (2018, April).Media-Rich Fake News Detection: A Survey. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 436-441). IEEE.

[6] Conroy, N. J., Rubin, V. L., & Chen, Y. (2015, November). Automatic deception detection: Methods for finding fake news. In Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community (p. 82). American Society for Information Science.

[7] A. Sinha, N. Arora, S. Singh, M. Cheema, and A. Nazir, "Fake Product Review Monitoring Using Opinion Mining," Torbet, Georgina. "U.S. Customers Spent over $6 Billion on Black Friday Purchases." Digital Trends, Digital Trends, 25 Nov. 2018, www.digitaltrends.com/web/shopping-totals-black-Friday/.

[8] M. Ott, C. Cardie, and J.T. Hancock. 2013. Negative Deceptive Opinion Spam. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

[9] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," Proceedings of EMNLP, pp. 79-86, 2002.

[10] Karami, Amir, and Bin Zhou. "Online review spam detection by new linguistic features." iConference 2015 Proceedings (2015).

[11] B. Pang and L. Lee, "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts," Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Article No. 271, 2004.

[12] C. Fellbaum, WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press, 1998.

[13] J. Kamps, M. Marx, R.J. Mokken, and M. Rijke, "Using WordNet to measure semantic orientations of adjectives," Proceedings of the Fourth International Conference on Language Resources and Evaluation, vol. IV, pp 1115-1118, 2004.

[14] https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

[15] https://docs.python.org/3.9/py-modindex.html

[16] M. Hu and B. Liu, "Mining and summarizing customer reviews," Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168-177, 2004.

[17] C. Castillo, M. Mendoza, and B. Poblete, “Predicting information credibility in time-sensitive social media,” Internet Research, vol. 23, no. 5, pp. 560–588, 2012.

[18] S. Jamshidi Nejad, F. Ahmadi-Abkenari and P. Bayat, "Opinion Spam Detection based on Supervised Sentiment Analysis Approach," 2020 10th International Conference on Computer and Knowledge Engineering (ICKE), 2020, pp. 209-214, DOI: 10.1109/ICCKE50421.2020.9303677.

[19] N. Jindal and B. Liu, "Analyzing and Detecting Review Spam," Seventh IEEE International Conference on Data Mining (ICDM 2007), 2007, pp. 547-552, DOI: 10.1109/ICDM.2007.68.

[20] Y. Lin, T. Zhu, H. Wu, J. Zhang, X. Wang, and A. Zhou, "Towards online anti-opinion spam: Spotting fake reviews from the review sequence," 2014

[21] IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), 2014, pp. 261-264, DOI: 10.1109/ASONAM.2014.6921594.

[22] Parikh, S. B., & Atrey, P. K. (2018, April).Media-Rich Fake News Detection: A Survey. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 436-441). IEEE.