Fake Review Detection using Principal Component Analysis and Active Learning

Faisal Muhammad Shah
Department of Computer Science & Engineering
Ahsanullah University of Science & Technology
Dhaka, Bangladesh

Sifat Ahmed
Department of Computer Science & Engineering
Ahsanullah University of Science & Technology
Dhaka, Bangladesh

ABSTRACT
E-commerce proved its importance based on the fact where time is the essence. People are relying on e-commerce more than before. With e-commerce comes a huge amount of user feedback based on the products they buy. As the internet has become cheaper and easy to get, more people are getting connected through different social media and platform where they are expressing product-related feedbacks. With the rise of e-commerce, people are relying more on product reviews to get a clear view and user experience. But there is no convincing way to authenticate the reviews posted on products on e-commerce websites. To generate more revenue and fulfill some immoral benefits, some sellers are making investments and hiring people to post fake reviews. These fake reviews are generated to convince people to buy the product. To detect these fake reviews, several methodologies were introduced. Most of the models are supervised models which rely on pseudo fake reviews or large scale labeled dataset. In this paper, a model has been proposed with a new technique which combines two different types of learning methods (active and supervised) by creating a manually labeled dataset. This model has 4 different filtering phases that are based on TF-IDF, Countvectorizer and n-gram features of the review content and then Principal Component Analysis to reduce the feature set. It achieves a very encouraging result while working on 2000 reviews from Amazon. In the best case precision, recall, and f-score are slightly above 91% and the accuracy achieved is up to 90%. After comparing the results with similar successful methods where PCA is used as a feature selection technique, it is quite clear that the proposed model is efficient and encouraging.

General Terms
Natural Language Processing, Spam review detection.

Keywords
Review spam detection, Fake review, PCA, Active Learning, Machine Learning.

1. INTRODUCTION
The usage of internet is increasing day by day as the world is becoming more digital and therefore the internet is easily accessible in both rural and urban areas. This has also brought commercial affairs to the web where not only the consumer but also the business is getting benefitted. People can easily post product reviews, views, experiences in blogs, discussion forums and on social platforms. These are addressed as user-generated contents. As people have the liberty to write whatever they want, there is no monitoring available. Sharing personal opinion, experience with a product is known as reviews. These reviews can attract and influence people to buy a product because people are getting a real-life experience on the product from someone else. These reviews have become a part and parcel to the buyers while buying a new product or an existing one. As these reviews make an influence on the buyers' side, some people provide fake reviews to increase the sale of the products found on e-commerce websites. These people are mainly known as opinion spammers and their activities are known as opinion spamming [11]. The number of fake reviews is increasing day by day. Some of the sellers are taking the chance to grow the business quickly by paying opinion spammers to write fake reviews. In fact, there are many websites that are paying to write fake reviews on different platforms [12]. Therefore detecting these fake reviews has become a serious issue to maintain the trust factor between the buyer and the customer.

There are many researchers who have come up with different spam detection techniques. But the major issue here is to find out an enriched real-life labeled dataset. Consequently, the existing solutions depend on pseudo fake reviews. Even some researchers are using a different psychological approach to detect a pattern of fake reviews and create a dataset. In [13] the authors used duplicate and nearly duplicate reviews as fake reviews. Some researchers used Amazon Mechanical Trunk (AMT) to write fake hotel reviews. But in [14], research proved that using pseudo-reviews as dataset might hamper the process of detecting spam reviews in a real-life environment. The authors also mentioned that these falsified reviews might show good results in experiments but in a real-life spam review detection their competence is doubtful. To overcome such problem, real-life Amazon reviews were used to create a spam review dataset using several approaches such as keyword-based search, sorting out reviews that are written for an exchange of a product, the number of reviews a single reviewer has written and their similarity measurements using Cosine Similarities also pseudo fake reviews written by experts. Our methodology is different from previous works for the following aspects-

- **Enriched and custom labeled dataset:** The dataset that has been prepared is labeled manually. From a chunk of 800,000 reviews collected from the dataset of the authors of [9] & [10], and sorted out based on review writers, writing pattern, keyword-based search such as "Honest review", "Fake review" and pseudo fake reviews written and combined them together to create an enriched dataset that adds more versatility and efficacy in real-life fake review detection.
- **Feature selection:** This paper includes Principal Component Analysis for content-based feature reduction such as tf-idf values, countvectorizer values, and n-gram values and some linguistic features which not only emphasizes the content of the reviews but also consider the context of it as accuracy was more concern here than efficiency.

2. RELATED WORKS

As per fake review detection is concerned, the supervised approach towards it is one of the most common methodologies used for detection. Although active learning is a new idea in this field, the authors of [3] have used active learning to create a hybrid dataset that can detect fake reviews from Yelp and Ott review dataset and performed very efficiently. Our model provides an extension of their work in fake review detection. The term fake review was first stated by [13] where they categorized it into different portions. After this, the authors of [15] tried to solve this problem by targeting individual spammers and their reviews at the same time the authors of [16] targeted the group of review spammers. In [17] and [18] other methods like time-series and distributional analysis were explored. The authors of [19] used a pattern recognition based on time-series where some other features were- rating deviation, content-based factors, and activeness of the reviewers. The authors of [4] tried something different here. They proposed a novel approach to detect fake reviews by applying a topic modeling method based on Latent Dirichlet Allocation. But the number of topics were fixed which were not enough while working with a real-life dataset. In [5], the authors proposed a BIRDNEST model that detects fake reviews based on time-series and user rating behavior to identify fraud review writers. In [6], the authors used the Convolutional Neural Network model to integrate the product-related review features through a product word composition model. The model did an impressive job with a specific feature set. At the same time, the dataset was created with real-life reviews. The authors of [7] defined a network schema with several review features such as review-behavioral, review-linguistic, user-linguistic, etc. to detect fake reviews posted on Yelp and Amazon. Most of the deceptive review detection approaches are based on supervised learning where the data were ad-hoc fake reviews on fabricated fake reviews [20] [21]. Manually deceptive fabricated reviews were not good enough to recreate the real-life problem environment and thus provided excellent results despite being a gold-standard dataset [22] [14]. Using the logistic regression learning model acquired Area Under the receiver operating characteristic curve (AUC) score of 0.78 [23]. The authors got this improved result after considering 24 features instead of only text features. The authors of [21] created a gold-standard dataset to identify review spam in which 3 groups of features were introduced- POS tag frequencies, LIWC features [24], n-grams. Naïve Bayes and SVM were trained and evaluated using 5-fold nested cross-validation. 89.8% accuracy was achieved by the model using bi-gram and LIWC features along with an SVM classifier. In [25] the authors conducted the research on the same dataset using n-gram features (unigram, bi-gram) with the SVM classifier which resulted in 86% accuracy. The authors of [8] showed how feature reduction can affect the sentiment analysis of online reviews. They used Principal Component Analysis on unigram features which increased the accuracy with SVM and Naïve Bayes classifier. One thing is clear that feature reduction on n-gram features could achieve better results in both positive and negative sentiment reviews. In [26], the authors used active learning along with clustering and random forest classifiers to detect email spam. This active learning method was also used in [27] to detect Wikipedia vandalism. Our target was to pick the best ideas from above and combine them to get the best possible output.

3. PROPOSED MODEL

In this section, the proposed model to detect review spam using hybrid machine learning technique has been elaborated. The whole methodology can be divided into five different phases- 1) Collecting fake and genuine reviews based on several criteria. 2) Preprocessing the collected data. 3) Creating a hybrid dataset with the help of active learning. 4) Reducing the feature set with PCA. 5) Supervised approach to detect fake reviews. The model is capable of handling both real-life and fabricated data as the training dataset contains real-life fake reviews and pseudo fake reviews. In previous researches, authors have followed their own approaches either by making own handmade dataset [28] or by crawling from different websites [29] which worked well in their environment and boundaries. In [22] the authors demonstrated that the model which was built for detecting review spams in [20] & [21] did not perform well at all on real-life data as the dataset was created based on the handcrafted deceptive reviews. To solve this problem a real-life and manually labeled dataset using different strategies has been used. By doing so, the efficiency increased more than expected in real-life reviews. The figure below illustrates the entire model-

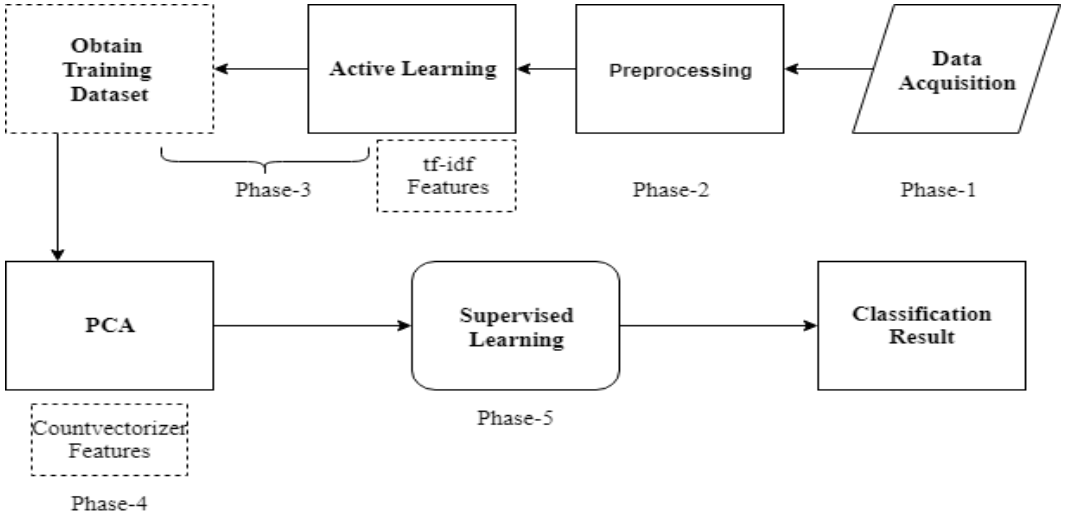

Fig 1: Proposed Model

In this model, the unlabeled data is taken as input and lightly preprocessed to remove noise and keep the necessary portions of the data. The preprocessed data then go to the active learner where training and test datasets are labeled. On completion, the learner starts to label the unlabeled dataset based on the accuracy measurement of the classification which ensures the quality of the training dataset. In this process SGD, Decision Tree and LinearSVC are used to classify the data. If the accuracy of classification is satisfactory then at first PCA is applied to reduce the feature set provided by count vectorizer using n-grams and then used to train a supervised model where several classifiers are used. Majority voting ensures the maximum accuracy of classification.

3.1 Data Acquisition

The first and foremost job was to get real-life reviews from Amazon. Raw Amazon reviews were collected from the authors of [9] & [10]. This dataset is known as "Amazon review data" which contains product reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996 – July 2014 (http://jmcauley.ucsd.edu/data/amazon/).

3.2 Dataset Creation and Preprocessing

The collected data was huge and totally raw. At the same time, PCA was applied for feature reduction. To help PCA work better, a dataset of the same categorical products (books) was needed to be collected. In the process, keyword-based search on a chunk of 8 million book reviews such as "Fake Review", "Honest Review", "False Information", "Paid Review" etc. were used. Based on their other activities such as rating, the time duration between reviews, word choice 2000 reviews were collected where 600 reviews are labeled as truthful reviews and 300 reviews are labeled as fake reviews. Leftover remained unlabeled. After preparing the data set, it was lightly preprocessed to deal with inconsistency and remove unnecessary and unwanted characters. At the same time stop words and punctuations were removed, slightly stemmed and lemmatized.

3.3 Applying the process

The whole experiment can be divided into two different phase. These phases are described below

PHASE-1: **Creating a dataset with active learning**

The first step of the whole approach is to construct a labeled dataset with the help of the active learning process which will be used later in phase-2. Active learning is a semi-supervised learning approach where a learning algorithm is used and the algorithm interactively queries the user to obtain the desired outputs at new data points [30] & [31]. The algorithm trains the model based on the training dataset and evaluates itself with the test dataset. After each evaluation, the algorithm selects some unlabeled data samples to classify by an expert. After classification, the newly labeled data samples are moved into the training set and the model starts to train itself again with the new training set and evaluates itself against the test dataset. The selection of unlabeled data which will be classified by an expert is selected based on the decision function which calculates the distance of the sample to the separating hyperplane. The range of this distance is from -1 to 1. To create a confidence level, the absolute value of the distance is used. Top six samples from either side with the highest and lowest confidence are selected for expert labeling. For feature selection, count vectorizer is used. These sparse vectors are fed to the classifiers. The classifiers used in this process are-

- Stochastic Gradient Descent Classifier (SGD)
- Linear SVC Classifier
- Decision Tree Classifier

Using this process, the unlabeled data were classified and the accuracy was checked after each iteration to maintain the quality of the dataset. If the above classifiers achieve an accuracy of more than 85% then it was considered that the model is performing well. After preparing the dataset, it was used in phase-2. After completing phase-1, there were 1125 truthful reviews and 875 fake reviews ready for phase-2.

PHASE-2: **Feature reduction using Principal Component Analysis.**

Principal Component Analysis (PCA) is applied to reduce the dimensions of inputs with larger dimensionality and the components are highly correlated. PCA creates a set of artificial variables which represents a set of the observed variable. The artificial variables are called principal components. These components are used criterion variable in other analysis. In PCA the components with the largest variation are chosen and the components with the least variation are eliminated.

Principal Component Analysis (PCA) was used to reduce the feature set in the experiment. PCA works better when applied to the same categorical products. Keeping this in mind, the dataset was created by selecting a specific product type (books). PCA constructs a new set of properties based on the combination of the old ones. Mathematically, PCA performs a linear transformation on the features by moving them to a new space composed of the principal components. Using this feature space, we are looking for properties that strongly differ from other classes. PCA looks for the properties that show as much variation across the classes as possible to build the principal component space. Components that have a higher variance were taken and discarded others for better accuracy and results.

PHASE-3: **The supervised approach for fake review detection.**

In this phase, supervised learning was used to detect fake reviews. After getting the dataset from phase-1 feature reduction algorithm was used on it. At this point, this data was fed into the supervised model to classify them accordingly. A classic but very efficient supervised model which is similar to several other successful supervised models [28] [21] [32] [33] [16] were used in the process.

Classifiers used

The classifiers used are-

- Naïve Bayes (NB)
- Support Vector Machine (SVC)
- Stochastic Gradient Descent (SGD)

The features collected from phase-2 are fed into these above classifiers. After getting the results, the best result was chosen out of a major voting system as done in [3].

4. EXPERIMENT

At this point, the process of creating the dataset has been discussed, classifiers used in this approach and the experimental setup in this section.

4.1 Dataset Collection and Preparation

In this section, the process of creating the dataset and preprocessing it has been discussed.

4.1.1 Labeled Data

In phase-1, raw amazon review data were collected from the authors of [9] & [10]. Before using this review data, it required cleansing, noise removal from the raw reviews to make them usable. After doing so, a keyword-based search strategy has been used to identify the writers of those reviews and other reviews written by the same writer. Based on their overall activities such as rating, the similarity between reviews, time duration 2000 were collected reviews where 900 reviews were labeled manually by us. Amongst them, 600 reviews are labeled as truthful reviews (ham) and 300 reviews are as fake reviews (spam). While choosing these 2000 reviews some fundamental verdicts were followed such as-

- The contents of the reviews must be in English.
- Reviews must not contain any hyperlinks, unnecessary and irrelevant character.
- Reviews must have a minimum length of 300 characters.

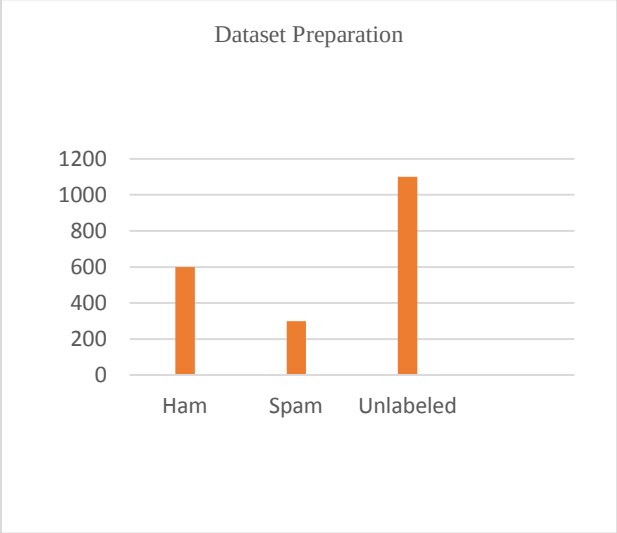

Fig 2: Dataset Splitting

4.1.2 Unlabeled Data

As mentioned earlier, 1100 data were kept unlabeled amongst the collected 2000 review data. These 1100 data contains both truthful reviews and fake reviews which are kept to be labeled by active learning.

4.1.3 Preprocessing

Before feeding the dataset into active learning, in phase-2, the dataset is lightly preprocessed to deal with inconsistency and unwanted characters. Stop words and punctuations were removed from the dataset as well as the dataset is slightly stemmed and lemmatized.

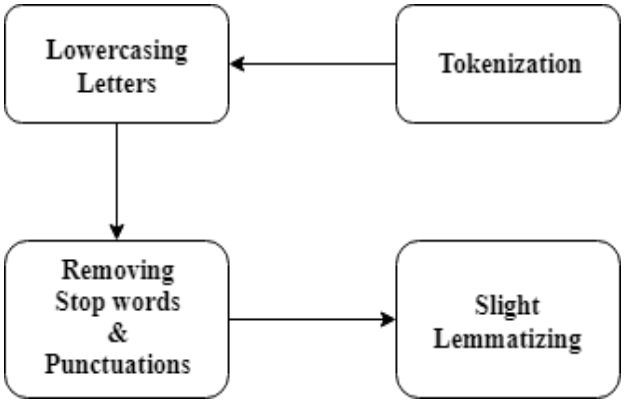

Fig 3: Preprocessing the reviews.

4.2.Data Labeling & Obtain Training Set

In phase-3, a dataset was created using active learning. At first 900 data were labeled manually where 600 are labeled as truthful reviews (ham) and 300 are labeled as fake reviews (spam). Amongst these 900 labeled reviews, 500 reviews are kept for the testing purpose and 400 reviews are used to train the active learning model initially. After that, remaining 1100 reviews were labeled with the help of active learning. After labeling, there are 1125 ham data and 875 spam data where 500 data were kept for testing purpose. The dataset is now ready with 1500 training data and 500 test data.

4.3 Evolution Metrics

The effectiveness of our model was evaluated on the basis of standard Accuracy (A), Precision (P), Recall (R), F1-score (F).

5. RESULT ANALYSIS

Now, in this section, evaluation and effectiveness of our proposed model on the basis of obtained results has been discussed. As mentioned before, after training the model with 400 labeled data, the learner labeled 1100 reviews. On completion, the dataset was fed into PCA for feature reduction. After reducing the features, the data were fed to the supervised model.

5.1 Classification Result

Two different strategies were used before feeding the data to classifiers. One is without feature reduction and one is with feature reduction applied. In both cases, the same classifiers were used to compare the effectiveness of feature reduction. The accuracy of all the classifiers with feature reduction was pleasing and better than without feature reduction. From Table I, it can be observed that the highest accuracy achieved is by Naïve Bayes and it is up to 85.5% and the precision is up to 87%. In this case, feature reduction isn't used before feeding it to classifiers. But in Table II, it can be seen that the accuracy has risen up to 90.1% and the precision is up to 93% with SVC.

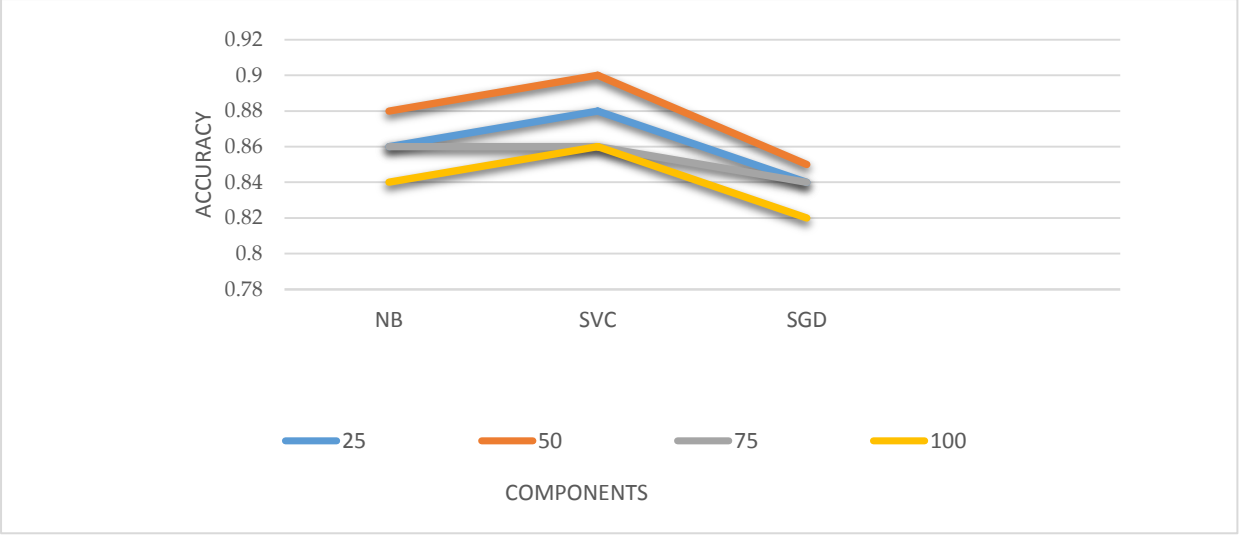

Fig 4: Accuracy of classifiers based on Principal Components selected

Table 1. Performance metric based on the number of selected principal components

Classifier	Components			
	25	50	75	100
NB	0.86	0.89	**0.86**	0.84
SVC	**0.88**	**0.90**	**0.86**	**0.86**
SGD	0.84	0.86	0.84	0.82

Table 2. Performance metrics for supervised learning without pca

Partition		Classifier	Accuracy	Precision	Recall	F1-Score
Training	1500	NB	**0.855**	**0.87**	**0.87**	**0.87**
		SVC	0.838	0.85	0.84	0.84
Test	500	SGD	0.842	0.86	0.85	0.85

Table 3. Performance metrics for supervised learning with pca

Partition		Components	Classifier	Accuracy	Precision	Recall	F1-Score
Training	1500	50	NB	0.89	**0.911**	0.88	0.902
			SVC	**0.90**	0.882	**0.92**	**0.931**
Test	500		SGD	0.86	0.900	**0.92**	0.920

The following table shows the accuracy and precision obtained from simulation with PCA-

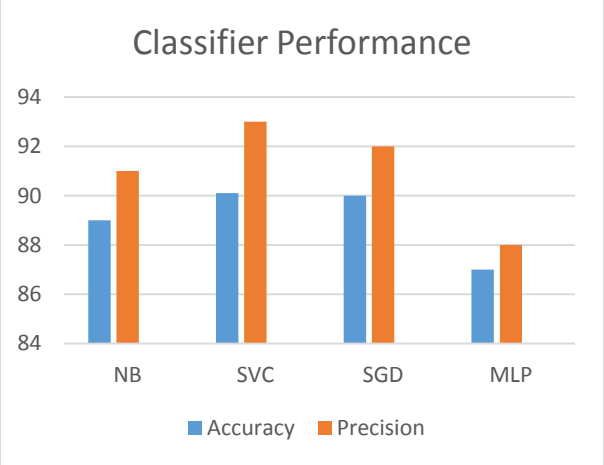

Fig 5: Accuracy & Precision achieved by Classifiers

From the above Table 2 and Table 3, it is very clear that PCA has a significant impact on fake review detection.

6. CONCLUSION

This paper proposes a hybrid method where active learning is used to label dataset and supervised classification method to identify review spam in Amazon review dataset. But using PCA as a feature selection method rather than using common feature selection methodology, has made an excellent increase in accuracy and precision.

In this experiment, 2000 amazon product reviews were used in which 900 were labeled by an expert and 1100 reviews were labeled by the active learning algorithm. After that, a labeled dataset was created where 1125 reviews were labeled as ham and 825 reviews were labeled as spam to conduct the experiment. In case of accuracy, Stochastic Gradient Descent outperformed all other classifiers by achieving the accuracy up to 90% and precision up to 91%. These results, when compared with some recent research which actually used PCA as a feature selection method for binary classification justifies the ability of our method. In the future, we look forward to using some new feature selection techniques and several tuning to our proposed model to achieve better results.

7. REFERENCES

[1] Li, H. et al., 2014. Spotting Fake reviews via Collective Positive Unlabeled Learning. 2014 IEEE International Conference on Data Mining, 18(3), pp.899-904. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7023420.

[2] Myle Ott, Claire Cardie, and Je_ Hancock. Estimating the prevalence of deception in online review communities. In Proceedings of the 21st International Conference on World Wide Web, pages 201-210. ACM, 2012.

[3] M.N. Istiaq Ahsan, Abdullah All Kafi and Tamzid Nahian. Faisal Muhammad Shah, "An Ensemble approach to detect Review Spam using hybrid Machine Learning Technique." 2016 19th International Conference on Computer and Information Technology (ICCIT).

[4] Kyungyup Daniel Lee, Kyungah Han and Sung-Hyon Myaeng. Capturing Word Choice Patterns with LDA for Fake Review Detection in Sentiment Analysis. WIMS 2016.Availableat:https://dl.acm.org/citation.cfm?id=2912868

[5] Bryan Hooi, Neil Shah, Alex Beutel, Stephan Gunnemann, Leman Akoglu, Mohit Kumar, Disha Makhija and Christos Faloutsos. "BIRDNEST: Bayesian Inference for Ratings-Fraud Detection". arXiv:1511.06030v2 [cs.AI] 2016.

[6] Chengai Sun et al. Chengai Sun, Qiaolin Du and Gang Tian. "Exploiting Product Related Review Features for Fake Review Detection". Available at: http://dx.doi.org/10.1155/2016/4935792

[7] Saeedreza Shehnepoor, Mostafa Salehi, Reza Farahbakhsh, Noel Crespi "NetSpam a Network-based Spam Detection Framework for Reviews in Online Social Media". arXiv: 1703.036009v1 [cs.SI] 10 Mar 2017.

[8] G.Vinodhini, RM.Chandrasekaran. "Effect of Feature reduction in Sentiment analysis of online reviews". ISSN:2278-1323; v2. IJARCET, 6 June, 2013.

[9] R. He, J. McAuley. "Ups and downs: Modeling the visual evolution of fashion trends with one class collaborative filtering". WWW, 2016.

[10] J. McAuley, C. Targett, J. Shi, A. van den Hengel. "Image-based recommendations on styles and substitutes". SIGIR, 2015.

[11] Algur, S., Hiremath, E., Patil, A. and Shivashankar, S., "Spam Detection of Customer Reviews from Web Pages." In Proceedings of the 2nd International Conference on IT and Business Intelligence.2010.

[12] Streitfeld, David. "Buy reviews on Yelp, get black mark." New York Times.Available: http://www.nytimes.com/2012/10/18/technology/yelp-tries-to-halt-deceptive-reviews. html. (2012)

[13] Jindal, Nitin, and Bing Liu. "Opinion spam and analysis." In Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 219-230. ACM, 2008.

[14] Mukherjee A, Venkataraman V, Liu B, Glance NS (2013) What yelp fake review filter might be doing? Boston, In ICWSM.

[15] Lim, Ee-Peng, et al. "Detecting product review spammers using rating behaviors." Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010.

[16] [Mukherjee, Arjun, Bing Liu, and Natalie Glance. "Spotting fake reviewer groups in consumer reviews." Proceedings of the 21st international conference on World Wide Web. ACM, 2012.

[17] Xie, Sihong, et al. "Review spam detection via temporal pattern discovery."Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012.

[18] Feng, S., Xing, L., Gogar, A., and Choi, Y. "Distributional Footprints of Deceptive Product Reviews". ICWSM. 2012

[19] Heydari, Atefeh, Mohammadali Tavakoli, and Naomie Salim. "Detection of fake opinions using time series." Expert Systems with Applications 58 (2016): 83-92.

[20] Ott M, Choi Y, Cardie C, Hancock JT (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-Volume 1 (pp. 309–319). Association for Computational Linguistics

[21] Ott M, Cardie C, Hancock JT (2013) Negative Deceptive Opinion Spam. In: HLT-NAACL., pp 497–501.

[22] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews," UIC-CS-03-2013. Tech. Rep., 2013.

[23] Jindal N, Liu B (2007) Review spam detection. In: Proceedings of the 16th international conference on World Wide Web (pp. 1189–1190). ACM, Lyon, France.

[24] Pennebaker, J.W. et al., The Development and Psychometric Properties of LIWC2007 The University of Texas at Austin. , pp.1–22.

[25] Heydari, A. et al., 2015. Detection of review spam: A survey. Expert Systems with Applications, 42(7), pp.3634–3642. Available at: http://dx.doi.org/10.1016/j.eswa.2014.12.029.

[26] DeBarr, Dave, and Harry Wechsler. "Spam detection using clustering, random forests, and active learning." Sixth Conference on Email and Anti-Spam. Mountain View, California. 2009.

[27] Chin, S.C., Street, W.N., Srinivasan, P. and Eichmann, D.,, April. "Detecting Wikipedia vandalism with active learning and statistical language models." In Proceedings of the 4th workshop on Information credibility (pp. 3-10). ACM.2010.

[28] Li, H. et al., 2014. Spotting Fake Reviews via Collective Positive-Unlabeled Learning. 2014 IEEE International Conference on Data Mining, 18(3), pp.899–904.Availableat:http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7023420.

[29] Myle Ott, Claire Cardie, and Je_ Hancock. Estimating the prevalence of deception in online review communities. In Proceedings of the 21st International Conference on World Wide Web, pages 201-210. ACM, 2012.

[30] Settles, Burr. "Active learning literature survey." University of Wisconsin, Madison 52.55-66 (2010): 11.

[31] Rubens, Neil, Dain Kaplan, and Masashi Sugiyama. "Active learning in recommender systems." Recommender systems handbook. Springer US,201 1. 735-767.

[32] B. Bigi, "Using Kullback-Leibler distance for text categorization," Proceeding ECIR'03 Proc. 25th Eur. Conf. IR Res., pp. 305–319, 2003.

[33] Li J, Ott M, Cardie C, Hovy E (2014) Towards a general rule for identifying deceptive opinion spam. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 1566–1576, Baltimore, Maryland, USA, June 23-25 2014. ACL.