Enhancing NLP based Text Mining Techniques for Fake Review Detection

Miss. Harsha.N.Dawda[1], Dr. S.W. Mohod[2]
[1,2]*Department of Computer Science and Engineering,*
Bapurao Deshmukh College of Engineering, Sevagram, Wardha

Abstract - Internet got boosted in few last years, people use web services more than expected. As the Internet got expanded people started their business online like E-Commerce websites. In today's modern era e-commerce has expanded broadly. Online Review has become an important factor for customers to buy products and judge how good the product is. However there are some users tend to give false reviews for the product, manufacturers and retailers are highly concerned with this customer feedback and review. This arise a serious concern that false reviewers devalue/value the products and the services. This is called Fake Review, where the users generate false review about the product for revenue gain/loss, hence fake review exist it is important to develop a technique to detect fake reviews. We have use active learning detection technique to detect the fake reviews.

Keywords - Amazon E-Commerce dataset, Dataset acquisition, Data pre-processing , Active Learning, Support Vector Machine, Perceptron Algorithm, Random Forest Algorithm.

I. INTRODUCTION

The Internet has changed our lives since it was introduced. With rapidly expansion and usage of Internet people are now totally dependent on the web services, which also changed people's behaviour of communicating and expressing their views. People post their views in their respective discussion groups, forums, social media, blogs and in e-commerce website for a product/service. These contents are user generated which are written in natural language. Opinion sharing on a product/service is based on their personal experience which is called as reviews.

In order to solve this malignant problem, we propose an interactive semi-supervised model to identify fake reviews which is evaluated later on using real life data and compared with some sophisticated prior research work. Active learning will be use to pre-process data set. Original dataset of Amazon reviews will be use to analyse and compare the results of different algorithm that are Support Vector Machine, Perceptron Algorithm, Random Forest Algorithm. In the literature, review spam has been categorized into three groups: Untruthful Reviews: The main concern of this paper. Reviews on Brands: Where the comments are only concerned with the brand or the seller of the product and fail to review the product. Non-Reviews: Those reviews that contain either unrelated text or advertisements. The first category, untruthful reviews, is of most concern as they undermine the integrity of the online review system.

Motivation - Online review is an important element in the era of E-commerce industry, where personal opinion on product is a convenient way to make decision whether to buy the product or not. That's why some people post poison reviews to harm the reputation of the respective product even though the product is good.

Writing fake review has become a serious issue, because of which user get false information about the product/service.

It's very serious review manipulation has been a problem for some time, and has only been growing.

It affects both sellers and buyers alike, negatively. It creates false depictions of inferior products; prevents better products from gaining traction and getting into the hands of people that need them; and of course limits your potential for sales as a result.

Companies that post fake reviews on websites to increase their ratings could face fines of thousands as part of a new government crackdown on misleading business practice.

Limitation - Sentiment analysis techniques used to extract and capture data for analysis in order to discern the subjective opinion of a document or collection of documents, like blog posts, reviews, news articles etc. For example: Twitter messages is challenging to mine because of the large and relevant sample information that they normally contain.

Effectively solving this task requires strategies that combine the small text content with prior knowledge and use more than just bag-of-words and it's usually important to look at data from the stand point of time because it changes over time according to a person's mood, world events.

II. LITERATURE REVIEW

M.N. Istiaq Ahsan, Tamzid Nahian, et.al have introduced an active learning approach to detect review spam using the TF-IDF features of the review content [16]. They have proposed an interactive semi-supervised model to identify fake reviews which is evaluated later on using real life data. They have not used the large-scale datasets from different domains in order to increase the size and diversity of the data to evaluate the heftiness of different classifiers. Diverse sets of tuning and smoothness techniques could be introduced. The feature set might be improved by using n-gram models (unigram, bigram, and trigram) with additional pre-processing techniques.

M.N. Istiaq Ahsan, Tamzid Nahian et .al, have introduced an ensemble learning approach which combines two different types of learning methods (active and supervised) by creating a hybrid dataset of both real-life and pseudo reviews [15]. This model holds 3 different filtering phases that is based on KL and JS distance, TF-IDF features and n-gram features of the review content. Large-scale datasets from different domains of different languages have not used in this system.

Elshrif Elmurngi, Abdelouahed Gherbi have analyse online movie reviews using SA methods in order to detect fake reviews [10]. Sentiment Analysis (SA) and text classification methods are applied to a dataset of movie reviews. The comparison of five supervised machine learning algorithms i.e Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbours (KNN-IBK), KStar (K*) and Decision Tree (DT-J48) for sentiment classification of reviews using two different datasets, including movie review dataset V2.0 and movie reviews dataset V1.0 has been given in this paper. The experiment shows that SVM algorithm outperforms other algorithms, and that it reaches the highest accuracy not only in text classification, but also in detecting fake reviews.

Lu zhang, Zhiangwu, Jiecao have proposed a partially supervised learning model (PSGD) to detect spammer groups. PSGD applies Positive Unlabelled Learning (PU-Learning) to study a classifier as spammer group detector from positive instances (labelled spammer groups) and unlabelled instances (unlabelled groups).They have extracted reliable negative set in terms of the positive instances and the distinctive features. By combining the positive instances, extracted negative instances and unlabelled instances, we convert the PU-Learning problem into the well-known semi-supervised learning problem, and then use Naive Bayesian model and EM algorithm to train a classifier for spammer group detection. They have implemented this system using Amazon.cn dataset [9].

ChiragVisani, NavjyotsinhJadeja, ManaliModi have constructed a feeling classifier, which can decide positive, negative and nonpartisan assumptions for a record. They have given the study about the algorithmic techniques for review spam discovery like Naive Bayes Classifier, Support Vector Machine (SVM), K-Nearest-Neighbour (Knn), Logistic Regression Classifier and detecting parameters like Group Time Stamp, Group Rating Fluctuation, Group Plagiarism, Cosine Similarity, Group Member Plagiarism, Early Time Stamp, Group Impact, Group Member Impact, Support Count, Review Length, Reviewer Investigation, and Stupidity. From the study we can say that support vector machine (SVM) outflank than the various administered strategies for survey spam identification [8].

Anna V. Sandifer, Casey Wilson, Aspen Olmsted have introduced a model for detecting fake online hotel reviews. They have extracted part-of-speech features from the data set and applied three classification techniques to identify fake online reviews. The three classification techniques are Multinomial Naive Bayes classifier, Bernoulli Naive Bayes classifier, and logistic regression classifier [7].

III. METHODS OVERVIEW

In this proposed system, there are five methods among this first four methods are about data sets and data pre-processing. Active Learning is a machine learning technique which is used to train datasets and used some classifier algorithms like Support Vector Machine (SVM), Random forest and Perceptron.

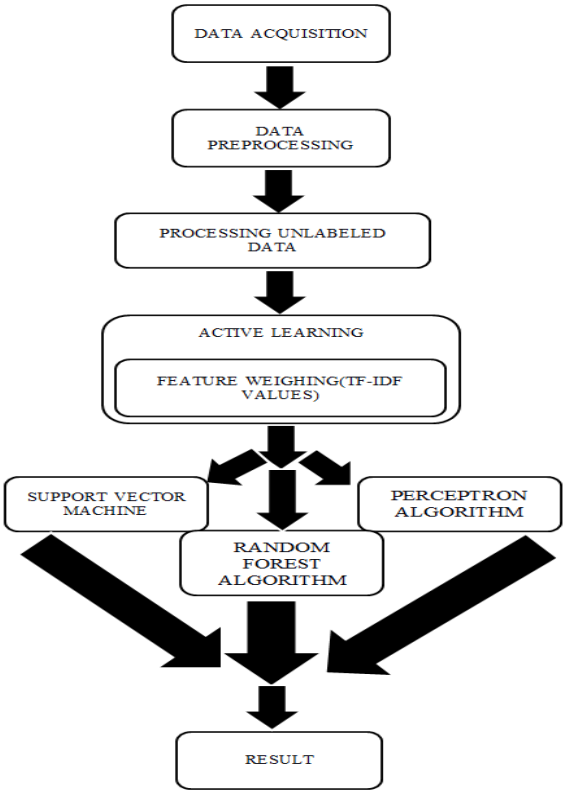

Figure 1: Data Flow Diagram of system

a) Data Acquisition - Data acquisitions are products and processes used to collect or analyze some phenomenon. We have used the original dataset of the Amazon reviews to test our methods of reviews classification and the dataset is used as unlabeled data. The Amazon dataset is used for both training and testing purpose in this method.

b) Data Preprocessing - Data pre-processing task is extremely crucial as it helps to generate organized information that is easy to understand and improves accuracy as well.

We apply a number of pre-processing techniques to deal with noisy, missing, and inconsistent data which might disturb decision-making process. Low-quality data will produce poor quality mining results and classification results. We ensure that the quality of the data we use in this experiment is up to the mark. Unstructured data in MS Excel format acquired from the source is converted into structured data i.e. in My SQL Database format.

Pre-processing procedures includes- tokenization & lowercasing letters, removing stop words, removing punctuations, stemming etc.

i). Part-of-Speech Tagging - also named grammatical tagging / word category disambiguation, is the technique of pattern up a word in a text (corpus) as matching to a specific part of speech, based on both its meaning and its context i.e. Its relationship with adjacent and corresponding words in a phrase, sentence, or paragraph. A basic form of this is frequently educated to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc.

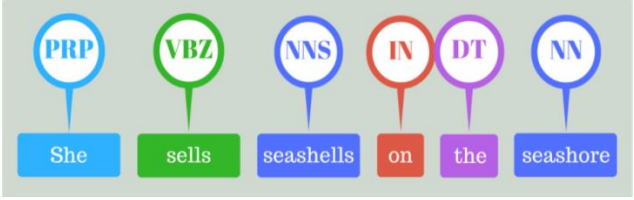

Figure 2: Diagram of POS tagging

ii). Dictionaries - We have maintained four dictionaries for data preprocessing

ii(a). Stemming - Stemming is the process of reducing a word to one or more stems. A *stemming dictionary* maps a word to its lemma (stem). A stemmer can use a stemming dictionary to improve the precision of a search. For example, the default stemming dictionary for English enables Mark Logic to map the words 'views', 'viewed', and 'viewing' back to their common stem, 'view'.

```
abet / abetting / abetted / abets / abettor / abetment
abort / abortion
abolish / abolition
abridge/abridges
abscond / absconding / absconds
accident / accidentally / accidents
accept / accepted / accepts / accepting
accord / accordingly / according / accords / accorded
accompany / accompanied
accuse / accused
achieve / achieved
act / acts / acting / action
```

Figure 3: Diagram of stemming words

ii(b). Stop word - Stop words are just a set of commonly used words in any language. Stop words are commonly eliminated from many text processing applications because these words can be distracting, non-informative. Text may contain stop words like 'the', 'is', 'and are'. Stop words can be filtered from the text to be processed.

ii(c). Lingo - In this dictionary we have maintained original words and shortcut words.

```
prilliant/briliant
location/locatn
about/abt
am/m
are/r
because/b'coz/cos/bcoz/
come/cm/cum
evening/eve/evng
and/N
morning/morn/morng
night/ni/nite/nit/nyt
please/plz/pls/plj
today/2day
what/wat/wht/wot/wt
```

Figure 4: Diagram lingo words

ii(d). SentiWordNet - We used this dictionary to calculate intensity and status of review. SentiWordNet dictionary calculates score to tagged words and score is given to Proposed SVM, perceptron, random forest to classify Reviews. Every word has positive and negative score already defined in the SentiWordNet dictionary so with help of that score, weighted score is assigned to tagged word to calculate its sentiment score

c) Processing Unstructured Data - The dataset have a delimiter and either fixed or variable width where the missing values are represented as blanks in between the delimiters. But sometimes we get data where the lines are not fixed width, image or pdf files. Such data is known as unstructured data. Processing unstructured data means tagging a label to the unprocessed data. In this step we make cluster head from the structured data.

d) Active Learning - Active learning is a special case of semi supervised machine learning which can interactively request the user or some sort of supervisor to determine the class of some unknown data points to achieve the desired results. We adopt this technique to train our model. This is a kind of a situation when unlabelled data is ample but labelling the whole dataset manually is extremely time consuming and labour intensive. So, the algorithm actively queries the user for labelling the new, confusing data points. In this type of learning, learner itself chooses the data point examples that's why it needs a much lower number of examples to learn a concept than it is required in typical supervised learning.

The algorithm trains the model based on a training dataset and evaluates using a test dataset. After each evaluation, the algorithm selects certain classifiers.

- **Classifiers - Support Vector Machine (SVM)** - SVM in machine learning is a supervised learning models with the related learning algorithm, which examines data and identifies patterns, which is used for regression and classification analysis.

Recently, many text-classification algorithms have been proposed, but SVM is still one of the most widely and most popular used classifiers. By applying the kernel equations arranges the data instances in such a way within the multi-dimensional space, that there is a hyper-plane that separates data instances of one kind into classes. The mostly used type of kernel equation is RBF. The function of kernel is to take data instance as input and transform it into the required output form. Once we manage to divide the data instances into two distinct categories, our aim is to get the best hyper-plane to separate the two types of instances.

The data points that were not linearly separable in the original domain have become linearly separable in the new domain, due to the application of a function (kernel) that transforms the position of the data instances from one domain to another. This is the basic idea for analyse large amount of data and patterns using Support Vector Machines and their kernel techniques. Whenever a new data instance

is encountered in the original domain, the same kernel function is applied to this new instance too, and its position in the new domain is found out.

- **Perceptron algorithm** - Perceptron is a machine learning linear classifier that helps to classified outcomes for computing.

Although it is suitable for large-scale learning we use this classifier to train our model. Some advantages that Perceptron classifier provides are learning rate is not required in this type of model, it is not regularized or penalized and this updates the model only when it commits a mistake that's why Perceptron is slightly faster to train with the hinge loss and that the resulting models are sparser.

- **Random Forest** - Random forest algorithm is a supervised classification algorithm. This algorithm generates the forest with a number of trees. In short, the **more trees in the forest** the more robust the forest looks like. Similarly, in the random forest classifier, the **higher the number** of trees in the forest gives **the high accuracy** results. Random forest algorithm can be state into two stages. (1) **Random forest** creation pseudo code. (2) **Pseudo code to perform prediction** from the created random forest classifier.

IV. IMPLEMENTATION

Step 1: Review Dataset

Here we have used Amazon review dataset. Electronic dataset is taken from amazon.com Download excel file for electronic dataset review which consist id, asins, name, rating, category and review. The review content must be in the English language.

Step 2: Preprocessing

Reviews contain information which are not clearly expressive or say meaning and need to be removed.

- Remove unwanted punctuations: All punctuations which are not necessary, it has been removed.
- Stop Word Removal: Some words used more and more time such words are called stop word. This pronouns, prepositions, conjunctions have no specific meaning.
- "i","a","an","is","are","as","at","from","in","this","on", " or","to","was","what","will","with" etc. are example of stop word, so these types of words has been vanished.
- Stemming: It converts word into its grammatical root form. Stemming technique converts word like "teach","teacher","teaching","teached","teaches" to root word teach.
- It minimizes the feature set and makes efficient classification performance by using java language. • Part of Speech Tagging:
- The Part-Of-Speech of a word is a linguistic category that is defined by its syntactic or morphological behavior. Noun, verb, adjective, adverb, pronoun, preposition, conjunction, and interjection are POS common categories.
- POS tagging is the task of marking each word in a sentence with its appropriate POS. we have used the Stanford tagger to tag the words. We have assign tag as verb, adjective, noun and adverb category.
- SentiWordNet dictionary calculates score to tagged words and score is given to Proposed SVM to classify Reviews. Every word has positive and negative score already defined in the SentiWordNet dictionary so with help of that score, weighted score is assigned to tagged word to calculate its sentiment score.

Step 3: Calculating Six parameters in database -

1. **Username** - checking whether name starts from alphabets or not if it is start from alphabets so we will set 1 in c_review column dynamically otherwise 0.
2. **Only star**: if user gives review so it will set 1 otherwise 0
3. **Status**: checking the status 1 indicates best, 2 indicates worst, 3 indicates good, 4 indicates bad, 5 indicates neutral.
4. **Intensity**: calculating the sentiment of review
5. **Review length**: calculating the length of review if review contains words >=5 then set 1 else 0
6. **Repeated review**: checking repeated review in text if review is repeated then it will set 1 otherwise 0
7. **Review Rating**: comparison of rating and sentiments.
8. **Text Comparison**: comparison of review and categories.
9. **Active Learning:** checking of total numbers of Parameter.
10. **Label**: if active learning column contains 6 so it will set correct otherwise incorrect.

Step 4: Implementing Algorithm -

Finally implement Svm, Perceptron and Random Forest Algorithm to find out total count of Genuine and fake reviews and their execution time.

V. CONCLUSION

This paper proposes an ensemble methodology for identifying Fake Review by renowned learning method (Active Learning) using real life data. System used several methods to analyse a dataset of Amazon product reviews. Had worked on sentiment classification algorithms to apply a supervised learning on electronic products of amazon reviews. Will experiment on approaches calculate the accuracy of perceptron, SVM and random forest i.e. sentiment classification algorithms. Additionally, system will able to classify how many given review are fake and genuine. Comparison will be done on the basis of results provided by different classification algorithm.

VI. REFERENCES

[1]. Xianguo Zhang, Xinyue Wang, Yang Liu, "Fake Reviews Detection Based on LDA", 4th IEEE International Conference on Information Management 2018.

[2]. DraskoRadovanovic, Bozo Krstajic, "Review Spam Detection using Machine Learning" 23rd International Scientific-Professional Conference on Information Technology (IT) 2018.

[3]. Xinyue Wang, Xianguo Zhang, Chengzhi Jiang, Haihang Liu, "Identification of Fake Reviews Using Semantic and

Behavioral Features" 4th IEEE International Conference on Information Management 2018.

[4]. Dhanya R, R KalaiSelvi , "A State of the Art Review on Copy Move Forgery Detection Techniques", Proceedings of 2017 IEEE International Conference on Circuits and Systems (ICCS 2017).

[5]. Huaxun Deng, Linfeng Zhao, NingLuo, Yuan Liu ,et al, "Semi-supervised Learning based Fake Review Detection", IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC) 2017.

[6]. SP.Rajamohana, Dr.K.Umamaheswari, "A Survey On Online Review Spam Detection Techniques", IEEE International Conference on Innovations in Green Energy and Healthcare Technologies(ICIGEHT'17).

[7]. Anna V. Sandifer, Casey Wilson , Aspen Olmsted, "Detection of fake online hotel reviews", The 12th International Conference for Internet Technology and Secured Transactions (ICITST-2017).

[8]. ChiragVisani ,Navjyotsinh Jadeja, ManaliModi, "A Study on Different Machine Learning Techniques for Spam Review Detection", International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017).

[9]. LU ZHANG, ZHIANG WU, AND JIE CAO, "Detecting Spammer Groups from Product Reviews: A Partially Supervised Learning Model", DOI10.1109/ACCESS.2017.2784370, IEEE Access.

[10]. ElshrifElmurngi, AbdelouahedGherbi, "Detecting Fake Reviews through Sentiment Analysis Using Machine Learning Techniques",DATA ANALYTICS 2017, The Sixth International Conference on Data Analytics.

[11]. ElshrifElmurngi, AbdelouahedGherbi, "An Empirical Study on Detecting Fake Reviews Using Machine Learning Techniques",The Seventh International Conference on Innovative Computing Technology (INTECH 2017).

[12]. Rajalaxmi Hegde1, Dr.Seema, "Aspect Based Feature Extraction and Sentiment Classification of Review Data sets using Incremental Machine learning Algorithm",3rd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB17).