DETECTION OF FAKE MOVIE REVIEWS USING DATAMINING TECHNIQUES

T.Sai Prasad Reddy1 M.Swathi2 M.Asritha3 K.Vinitha4 G.Priyanka5
Associate Professor1, U.G. Scholar2,3,4,5 Dept of Computer Science and Engineering, Narayana Engineering College, Nellore, India.

Abstract: Reviews are having high impact on bussiness. Deciding for purchase of things mostly depends on reviews given by the users. Hence, some individuals or groups attempt to manipulate reviews for his or her own interests. therefore the customers attracts to those fake reviews easily and therefore the sales for particular business could also be increased. This paper introduces some semi-supervised and supervised text mining models to detect fake movie reviews. Some approaches are review content based and a few are supported behavior of the user who is posting reviews. Content based study focuses on what's written on the review that's the text of the review where user behavior based method focuses on country, ip-address, number of posts of the reviewer etc. Here we use three techniques called genre identification, detection of behavioral deception and text categorization. By using these features we will reduce over fitting and obtain the very best accuracy by using supervised classification with Naive Bayes classifier.

Keywords: SVM , Feature Extraction, Supervised Learning .

I.INTRODUCTION

Data mining could be a field of research that has emerged within the 1990s, and is extremely popular today, sometimes under different names like “big data” and “data science“, which have an analogous meaning. to allow a brief definition of information mining, it will be defined as a collection of techniques for automatically analyzing data to find interesting knowledge or pasterns within the data. The reasons why data processing has become popular is that storing data electronically has become the bottom which transferring data can now be done very quickly because of the fast computer networks that we've today. Thus, many organizations now have huge amounts of information stored in databases, that must be analyzed. Having lots of information in databases is great. However, to essentially get pleasure from this data, it's necessary to investigate the information to grasp it. The data which is not understandable are useless. So a way to analyze the information stored in large databases? Traditionally, data has been analyzed by hand to find interesting knowledge. However, this can be time-consuming, vulnerable to error, doing this might miss some important information, and it's just not realistic to try to to this on large databases. to handle this problem, automatic techniques are designed to investigate data and extract interesting

patterns, trends or other useful information. this can be the aim of information mining. To perform data processing, a process consisting of seven steps is typically followed. This term is known as the "Knowledge Discovery in Database" (KDD) process. 1. Data cleaning: This step consists of cleaning the information by removing noise or other inconsistencies that might be an issue for analyzing the information. 2. Data integration: This step consists of integrating data from various sources to organize the information that must be analyzed. as an example, if the information is stored in multiple databases or file, it's going to be necessary to integrate the information into one file or database to investigate it. 3. Data selection: This step consists of choosing the relevant data for the analysis to be performed. 4. Data transformation: This step consists of remodeling the information to a correct format that may be analyzed using data processing techniques. as an example, some data processing techniques require that every one numerical values are normalized. 5. Data mining: This step consists of applying some data processing techniques (algorithms) to investigate the information and see interesting patterns or extract interesting knowledge from this data. 6. Evaluating the knowledge that has been discovered: This step consists of evaluating the knowledge that has been extracted from the information. this may be worn out terms of objective and/or subjective measures. 7. Visualization: Finally, the last step is to visualise the knowledge that has been extracted from the information. Some approaches are review content based and a few are supported behavior of the user who is posting reviews. Content based study focuses on what's written on the review that's the text of the review where user behavior based method focuses on country, ip-address, number of posts of the reviewer etc. Most of the proposed approaches are supervised classification models. Few researchers, even have worked with semi-supervised models. Semi-supervised methods are being introduced for lack of reliable labeling of the reviews. In this paper, we make some classification approaches for detecting fake online reviews, a number of which are semi supervised et al are supervised. For semi-supervised learning, we use Expectation-maximization algorithm. Statistical Naive Bayes classifier and Support Vector Machines(SVM) are used as classifiers in our research work to boost the performance of classification. we've mainly focused on the content of the review based approaches. As feature we've used word frequency count, sentiment polarity and length of review.

II.RELATED WORK

A number of studies are conducted which focused on spam detection in e-mail and on the net , however, only recently have any studies been conducted on opinion spam. For detecting

fake reviews and located that opinion spam is widespread and different in nature from either e-mail or Web spam. they need classified spam reviews into 3 types: Type 1, Type 2 and kind 3. Here Type 1 spam reviews are untruthful opinions that commit to mislead readers or opinion mining systems by giving untruthful reviews to some target objects for his or her own gains. Type 2 spam reviews are brand only reviews, those who comment only on the brand and not on the products. Type 3 spam reviews aren't actually reviews, they're mainly either advertisements or irrelevant reviews which don't contain any opinions about the target object or brand. Although humans detect this type of opinion spam they have to be filtered, because it's visiting be a nuisance for the very best user. Their investigation was supported 5.8 million reviews and some of.14 million reviewers (members who wrote a minimum of 1 review) crawled from amazon.com which they have discovered that spam activities are widespread. they need regarded spam detection as a classification problem with two classes, spam and non-spam. And have built machine-learning models to classify a review as either fraud or not. they need detected type 2 and kind 3 spam reviews by using supervised learning with manually labelled training examples and located that the highly effective model is logistic regression model. However, to detect type 1 opinion spam, they might have had to manually label training examples. Thus that they had to use duplicate fake reviews as positive training examples and other reviews as negative examples to create a model.

In the paper "Finding Deceptive Opinion Spam by Any Stretch of the Imagination" they need given focus to the deceptive opinion fraud i.e. the fictional opinions which are deliberately written to sound authentic so on deceive the user. The user cannot easily identify this type of opinion spam. They need mined all reviews for 20 most famous hotels in Chicago area and fake opinions were gathered for the identical hotels. They first asked human judges to guage the review then they need automated the task for the identical set of reviews, which they found that automated classifiers outperform humans for every metric. The performance was compared with the psycholinguistic fraud detection and genre identification which were outperformed by n-gram based text categorization, but a combined classifier of n-gram and psychological deception features achieved 90% accuracy. Authors have found that a mixture of linguistic and behavioral features comparatively yielded more accuracy.

Several data preprocessing steps are performed on the above dataset before it's used. • Removal of anonymous users: We first remove anonymous users and their reviews. Each anonymous user id could even be used by multiple persons.

• Removal of duplicate : We also identify sets of duplicates within the dataset and deduct

them apart from one representative one per set. This step is significant since Amazon.com maintains duplicate products (essentially the identical product with some very minor variations, e.g. color) and replicates reviews across them. In other words, given a gaggle of duplicate products, a review written on a product have gotten to be replicated and added to other products during this set. Using the identical reviews, we detect sets of such products and randomly choose one representative product from each set to stay while removing others.

• Removal of inactive users and unpopular products: To specialise in users who are active and products that attract some user attention, our dataset includes only users and products with no fewer than 3 reviews. This is often done by repetitively applying minimum number of reviews threshold onusers and products in alternate order until all users and products meet the sting .

• Resolution of name name name name synonyms: We acknowledged that the products' brand names suffer from the synonymy problem which involves multiple brand names assigned to the identical brands. E.g., the brand "HP" may even be called "Hewlett Packard" or "HP Technology". Fortunately, there are only few many brand names in Products. We therefore were ready to resolve synonyms by manual inspection and replace them by the representative brand names.

Disadvantages

 within the prevailing work, the system uses only to semi-supervised learning.

III.PROPOSED WORK

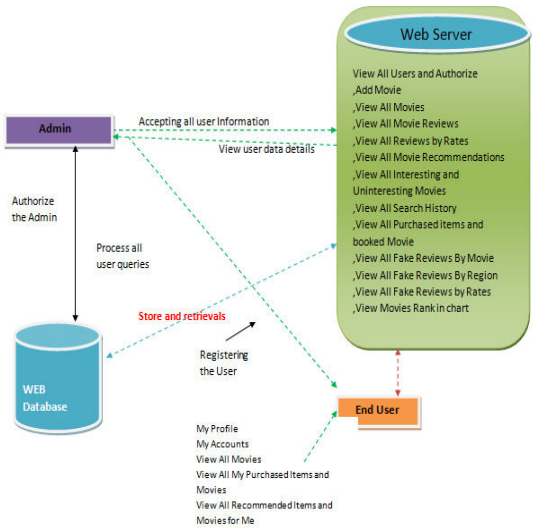
The following feature points were chosen to be extracted and used for the experiments from the dataset:

• Sentiment Polarity

• Parts of Speech (POS) tags

• Linguistic Inquiry and Word Count (LIWC)

• Bigram frequency counts



According to the observations, fake reviews have more positive/ negative sentiment than the normal ones generated by actual customers. A specific product would be described by some special feature words and mawkish words when the spammers write the fake reviews. as an example, product features within the movie domain a touch just like the name of the movie and mawkish words like "extremely comfortable" are widely used. In other domains,smart phone is often evaluated by

"sleek" and "stable" and keyboard by "wireless" and "mechanical." This product oriented data affects the performance of the prediction; thus integrating it into a classification model will benefit the classifier lots. For identification of phony surveys, we start with crude content information. we've utilized a dataset which was at that time named by the past specialists. We evacuate pointless writings like article and relational words within the knowledge. At that time these content information are changed over into numeric information for creating them appropriate for the classifier. Significant and vital highlights are separated and afterward classification process occurred.

The process of detecting the fake review is:

1) Each review goes through tokenization process first. Then, unnecessary words are removed and candidate feature words are generated.

2) Each candidate feature words are checked against the dictionary and if it's entry is obtainable within the dictionary then it's frequency is counted and added to the column within the feature vector.

3) Alongside with counting frequency, The length of the review is measured and added to the feature vector.

4) Finally, sentiment score which is obtainable within the knowledge set is added within the feature vector. we've assigned negative sentiment as zero valued and positive sentiment as some positive valued within the feature vector. For detecting the fake reviews we used the expectation-maximization algorithm(EM).As classifier, we've used Support Vector machines(SVM) and Naive Bayes(NB) classifier with EM algorithm.

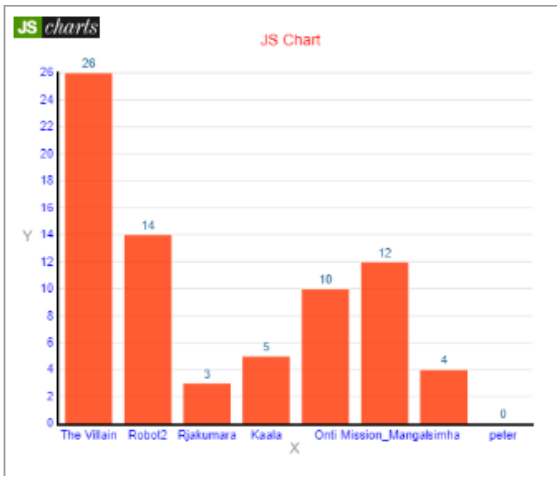
IV.PROPOSED ALGORITHM

The Expectation Maximization algorithmis designed to label unlabeled data to be used for training. The algorithm is operated as: a classifier is derived from the labeled dataset. This classifier is then accustomed label the unlabeled dataset. Let this predicted set of labels be PU. Now, another classifier springs from the combined sets of both labeled and unlabeled datasets and is used to classify the unlabeled dataset again. This process is repeated until the set PU get stabilized. After a stable PU set is produced, we learn the classification algorithm with the combined training set of both labelled and unlabeled datasets and deploy it for predicting test dataset. The learning of the algorithm with the conjunction of the labeled and predicted labeled sets is the Expectation step and so the prediction of the labels of the unlabeled set is the Maximization step.

Support Vector Machine (SVM) could also be a supervised machine learning algorithm which could be used for classification or regression problems. It uses the kernel trick to transform your data then supported these transformations it finds an optimal boundary between the outputs. Simply put, it does some

extremely complex data transformations, then figures out some way to seperate your data supported the labels or outputs you've defined.

Naive Bayes is that the straightforward and fast classification algorithm, which is suitable for an oversized chunk of data. Naive Bayes classifier is utilized in various applications like spam detection, text classification, sentiment analysis, and recommender systems. It uses Bayes theorem of probability for prediction of unknown class.Whenever you perform classification, the first step is to grasp the matter and identify potential features and label. As an example, within the case of a loan distribution, bank manager's identify customer's occupation, income, age, location, previous loan history, transaction history, andcredit score. These characteristics are called features which help the model classify customers. The classification has two phases, a learning phase, and so the evaluation phase. within the educational phase, classifier trains its model on a given dataset and within the evaluation phase, it tests the classifier performance. Performance is evaluated on the thought of various parameters like accuracy, error, precision, and recall. By this we get the easiest accuracy and so the graph is shown as below:



Advantages:

- The system is extremely fast and effective thanks to semi-supervised and supervised learning.
- Focused on the content of the review based approaches. As feature we have used word frequency count, sentiment polarity and length of review.

V.CONCLUSION

By using the expectation maximization algorithm, Naïve bayes classifier and support vector machine the performance has been improved and gained the highest accuracyby the usage of the semi supervised learning and supervised learning. In future, as a enhancement we can use this for the product related data and can provide the accuracy.

VI.REFERENCES

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

[2] A. Heydari, M. A. Tavakoli, N. Salim, and Z. Heydari, ”Detection of review spam: a survey”, Expert Systems with Applications, vol. 42, no. 7, pp. 3634–3642, 2015.

[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), vol. 1, pp. 309–319, Association for Computational Linguistic Portland, Ore, USA, June 2011.

[4] J. W. Pennebaker, M. E. Francis, and R. J. Booth, ”Linguistic Inquiry and Word Count: Liwc,” vol. 71, 2001.

[5] S. Feng, R. Banerjee, and Y. Choi, “Syntactic stylometry for deception detection,”in Proceeding of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers, Vol. 2, 2012.

[6] 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), 2014.

[7] 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.

[8] 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.

[9] J. Karimpour, A. A. Noroozi, and S. Alizadeh, “Web spam detection by learning from small labeled samples,” International Journal of Computer Applications, vol. 50, no. 21, pp. 1–5, July 2012.