DETECTION OF FAKE MOVIE REVIEWS USING DATA MINING TECHNIQUES

Dr A.Srilakshmi, Lecturer in Computer Applications,
Govt Degree College,Koduru
M. Swathi, M. Asritha, K. Vinitha, G. Priyanka
Narayana Engineering College, Nellore

Abstract: Reviews have great impact on today's business and commerce. Decision making for purchase of things mostly depends on reviews given by the users. Hence, some individuals or groups try to manipulate reviews for their own interests. So the customers attracts to these fake reviews easily and the sales for particular business may 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 some are based on behavior of the user who is posting reviews. Content based study focuses on what is written on the review that is 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 can reduce over fitting and get the highest accuracy by using supervised classification with Naive Bayes classifier.

Keywords: SVM , Feature Extraction, Supervised Learning .

I.INTRODUCTION

Data mining is a field of research that has emerged in the 1990s, and is very popular today, sometimes under different names such as "big data" and "data science", which have a similar meaning. To give a short definition of data mining, it can be defined as a set of techniques for automatically analyzing data to discover interesting knowledge or pasterns in the data.

The reasons why data mining has become popular is that storing data electronically has become very cheap and that transferring data can now be done very quickly thanks to the fast computer networks that we have today. Thus, many organizations now have huge amounts of data stored in databases, that needs to be analyzed.

Having a lot of data in databases is great. However, to really benefit from this data, it is necessary to analyze the data to understand it. Having data that we cannot understand or draw meaningful conclusions from it is useless. So how to analyze the data stored in large databases? Traditionally, data has been analyzed by

hand to discover interesting knowledge. However, this is time-consuming, prone to error, doing this may miss some important information, and it is just not realistic to do this on large databases. To address this problem, automatic techniques have been designed to analyze data and extract interesting patterns**,** trends or other useful information. This is the purpose of data mining**.**

To perform data mining, a process consisting of seven steps is usually followed. This process is often called the "Knowledge Discovery in Database" (KDD) process.

1. Data cleaning: This step consists of cleaning the data by removing noise or other inconsistencies that could be a problem for analyzing the data.
2. Data integration: This step consists of integrating data from various sources to prepare the data that needs to be analyzed. For example, if the data is stored in multiple databases or file, it may be necessary to integrate the data into a single file or database to analyze it.
3. Data selection: This step consists of selecting the relevant data for the analysis to be performed.
4. Data transformation: This step consists of transforming the data to a proper format that can be analyzed using data mining techniques. For example, some data mining techniques require that all numerical values are normalized.
5. Data mining: This step consists of applying some data mining techniques (algorithms) to analyze the data and discover 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 data. This can be done in terms of objective and/or subjective measures.
7. Visualization: Finally, the last step is to visualize the knowledge that has been extracted from the data.

Some approaches are review content based and some are based on behavior of the user who is posting reviews. Content based study focuses on what is written on the review that is 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, also 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, some of which are semi supervised and others 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 improve the performance of classification. We have mainly focused on the content of the review based approaches. As feature we have used word frequency count, sentiment polarity and length of review.

II.RELATED WORK

A number of studies have been conducted which focused on spam detection in e-mail and on the web, however, only recently have any studies been conducted on opinion spam. For detecting fake reviews and found that opinion spam is widespread and different in nature from either e-mail or Web spam. They have classified spam reviews into 3 types: Type 1, Type 2 and Type 3. Here Type 1 spam reviews are untruthful opinions that try to mislead readers or opinion mining systems by giving untruthful reviews to some target objects for their own gains. Type 2 spam reviews are brand only reviews, those that comment only on the brand and not on the products. Type 3 spam reviews are not actually reviews, they are mainly either advertisements or irrelevant reviews which do not contain any opinions about the target object or brand. Although humans detect this kind of opinion spam they need to be filtered, as it is a nuisance for the end user. Their investigation was based on 5.8 million reviews and 2.14 million reviewers (members who wrote at least one review) crawled from amazon.com and they have discovered that spam activities are widespread. They have 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 spam or non-spam. They have detected type 2 and type 3 spam reviews by using supervised learning with manually labelled training examples and found that the highly effective model is logistic regression model. However, to detect type 1 opinion spam, they would have had to manually label training examples. Thus they had to use duplicate spam reviews as positive training examples and other reviews as negative examples to build a model.

In the paper "Finding Deceptive Opinion Spam by Any Stretch of the Imagination" by Ott, et al. 2011, they have given focus to the deceptive opinion spam i.e. the fictitious opinions which are deliberately written to sound authentic so as to deceive the user. The user cannot easily identify this kind of opinion spam. They have mined all 5-star truthful reviews for 20 most famous hotels in Chicago area

from trip advisor and deceptive opinions were gathered for the same hotels using amazon mechanical trunk (AMT). They first asked human judges to evaluate the review and then they have automated the task for the same set of reviews, and they found that automated classifiers outperform humans for each metric. The task was viewed as standard text categorization task, psycholinguistic deceptive detection and genre identification. The performance from each approach was compared and they found that the psycholinguistic deceptive detection and genre identification approach was outperformed by n-gram based text categorization, but a combined classifier of n-gram and psychological deception features achieved nearly 90% cross-validated accuracy. Finally they came into a conclusion that detecting deceptive opinions is well beyond the capabilities of humans. Since then, various dimensions have been explored: detecting individual (Lim et al., 2010)[6] and group spammers (Mukherjee et al., 2012), time-series (Xie et al., 2012)[8] and distributional analysis (Feng et al., 2012a) . Yoo and Gretzel (2009) gather 40 truthful and 42 deceptive hotel reviews and, using a standard statistical test, they have manually compared the psychologically relevant linguistic differences between them. In (Mukherjee, et al., 2013), authors have briefly analyzed "What yelp filter might be doing?" by working with different combination of linguistic features like unigram, bigram, distribution of parts of speech tags and yielding detection accuracy. Authors have found that a combination of linguistic and behavioral features comparatively yielded more accuracy.

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

• Removal of duplicate : We also identify sets of duplicates in the dataset and remove them except for one representative one per set. This step is necessary since Amazon.com maintains duplicate products (essentially the same product with some very minor variations, e.g. color) and replicates reviews across them. In other words, given a set of duplicate products, a review written on a product will be replicated and added to other products in this set. Using the identical reviews, we detect sets of such products and randomly choose one representative product from each set to keep while removing others.

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

• Resolution of brand name synonyms: We found out that the products' brand names suffer from the synonymy problem which involves multiple brand names assigned to the same brands. E.g., the brand"HP"may also be called "Hewlett Packard" or "HP Technology". Brand synonyms prevent us from grouping products by their brands. Fortunately, there are only few hundreds of brand names in MProducts. We therefore were able to resolve synonyms by manual inspection and replace them by the representative brand names.

Disadvantages

- In the existing work, the system uses only to semi-supervised learning.
- Only Text Classification as sentiment text and it never finds fake review.

III.PROPOSED WORK

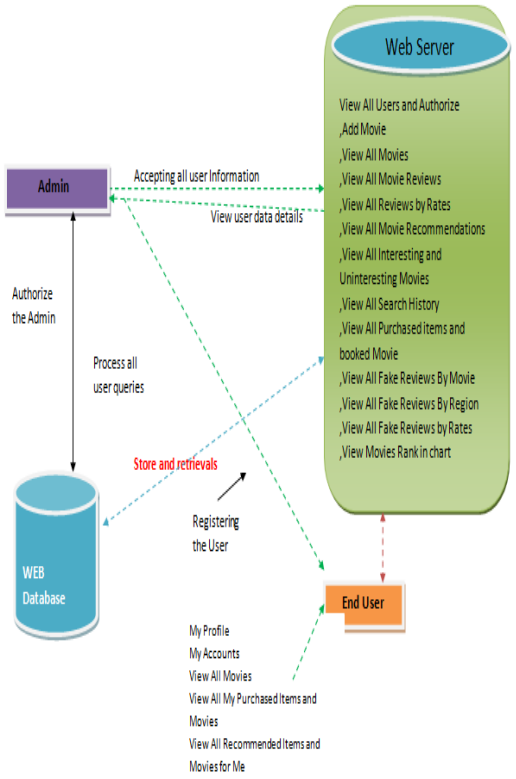
The following feature points were chosen to be extracted and used for the experiments from the dataset:

• 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. That is, review spammers emphasized some features using more positive/ negative words to agitate for/slander a product. This means that a particular product would be described by some special feature words and sentimental words when the spammers

write the fake reviews. For example, product features in the movie domain like the name of the movie and sentimental words like“extremely comfortable”are widely used. In other domains, according to their findings ,smartphone is often evaluated by “sleek” and “stable”and keyboard is evaluated by “wireless” and “mechanical.” This product oriented information affects the performance of the prediction; thus integrating it into a classification model will benefit the classifier a lot. For identification of phony surveys, we start with crude content information. We have utilized a dataset which was at that point named by the past specialists. We evacuate pointless writings like article and relational words in the information. At that point these content information are changed over into numeric information for making 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 available in the dictionary then it’s frequency is counted and added to the column in the feature vector that corresponds the numeric map of the word.

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

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

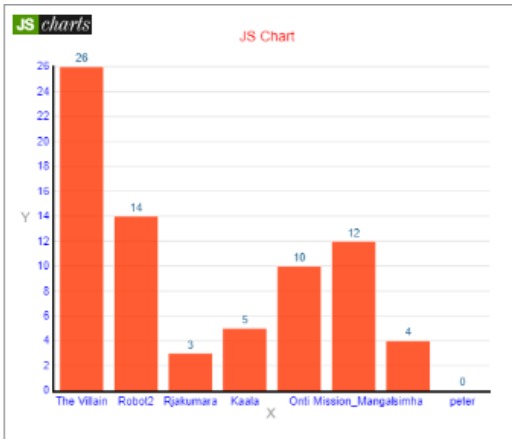
IV.PROPOSED ALGORITHM

The Expectation Maximization algorithmis designed to label unlabeled data to be used for training. The algorithm operates as follows: a classifier is first derived from the labeled dataset. This classifier is then used to label the unlabeled dataset. Let this predicted set of labels be PU. Now, another classifier is derived 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 stabilizes. 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. Here ,the learning of the algorithm

with the conjunction of the labeled and predicted labeled sets is the Expectation step (E-step) and the prediction of the labels of the unlabeled set is the Maximization step (M-step).

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs. Simply put, it does some extremely complex data transformations, then figures out how to seperate your data based on the labels or outputs you've defined.

Naive Bayes is the most straightforward and fast classification algorithm, which is suitable for a large chunk of data. Naive Bayes classifier is successfully used in various applications such as spam filtering, 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 understand the problem and identify potential features and label. Features are those characteristics or attributes which affect the results of the label. For example, in 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 known as features which help the model classify customers. The classification has two phases, a learning phase, and the evaluation phase. In the learning phase, classifier trains its model on a given dataset and in the evaluation phase, it tests the classifier performance. Performance is evaluated on the basis of various parameters such as accuracy, error, precision, and recall. By this we get the highest accuracy and the graph is shown as below:



Advantages:

- The system is very fast and effective due 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 deceptiondetection,"inProceedings 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.