International Journal For Technological Research In Engineering
Volume 6, Issue 9, May-2019 *ISSN (Online): 2347 - 4718*

ANALYSIS OF CLASSIFIERS FOR FAKE REVIEW DETECTION

Ms. Rajshri P. Kashti[1], Dr. Prakash S. Prasad[2]
[1]M. Tech Scholar, [2]Head of Department
[1]Department of CSE, Priyadarshini Institute of Engineering& Technology Nagpur, India

Abstract: World of internet has a great impact on online shopping as new buyers use experience of others about the product or service. Opinion or review comment given by someone in order to ruin the reliability of the product or service is considered as Fake review or spam.Thus it is an extremely important to verify their reliability before buying product. Natural language processing techniques are widely used for spam detection. Different NLP techniques are proposed earlier for detecting the review spam such as active learning, n-gram patterns etc. This paper presents an active learning method using different classification algorithm for detecting fake and genuine reviews. The classification algorithms suggested and implemented are rough-set, decision tree, random forest and support vector machine. This paper studies the effectiveness of algorithm used for fake review detection.

Index Terms: Active learning, Decision tree algorithm, Fake reviews, Random Forest, Rough Set Classifier, support vector machine.

I. INTRODUCTION

Review can consists of various terms like category of product to which it belongs to or what features the buyers like most and so he/she uses words to express the feeling like good, bad,excellent about product or service. Now here another factor comes into play and that is about the spammers. Spammer is the individual who tries to produce false impression on another buyer by giving wrong comment. Now the question arises how can we identify the given review is spam or fake. Since the review is in textual format, natural language processing techniques are used.

Natural language processing are commonly used for Text mining techniques. Review comments are written in natural English language. So it is important to preprocess them before they can be preceded for classification. Various factors can be used for detecting fake and genuine reviews. As database of review can consists of too many factors, it's extremely important to choose those factors wisely and it must be feasible to use them for detecting fake and genuine. Thus our paper is suggests the methods and algorithms for detection of fake and genuine review techniques accuracy for amazon dataset. This paper evaluates the accuracy of classifiers namely Roughset, SVM and Random forest & decision tree.

II. FACTORS CONSIDERED FOR FAKE REVIEW DETECTION

Review Content:

Lexical features such as part-of-speech, and other lexical attributes. Content and text similarity of reviews from different reviewers.

Semantic inconsistency

Product Related Features: E.g., product description based on its category.

Numeric Username: genuine customer must have his/her name and it should not be numeric. Username with address in alphanumeric notation indicate genuine user.

Star (*) Rating Only: Sometimes customer just put Star Rating for the product or service. But if he is really want to say good or bad about product , he does give review comment as well.

Rating Vs Review Sentiment: Here sentiment value is calculated for review based on threshold set for "excellent", "good", "bad", " worse" etc. If the sentiment calculated for the review does not match to the rating, given reviews can be tagged as fake.

Review Length: here we consider that the length of review must be at least 5 words which should comprise features of product or services.

III. CLASSIFIERS

3.1 Rough Set Classifier

Rough set classifier makes use of decision rules for training our model. Rough set classifier produces minimal decision rules that makes use of union and intersection rules from set theory (AND-OR construct in programming) for review parameter. It is a hybridized tool that includes sequence Arithmetic, Rough Set Theory and Concept Lattice[14]. The accuracy level of this classifier is 97.7%. Thus this model requires extra time and space for further classification of the output sequence into fake or genuine. A classification algorithm repeatedly use the knowledge acquired in the previous situation to make forecast in new situation. Thus it helps to classify the object (in this case reviews) that has not seen earlier. Each new review is assigned to a class (two classes here specifically fake and genuine)belonging to a predefined set of classes on the basis of observed values of suitably chosen attributes (features). Rough set classifiers are used for generating minimal set of rules that helps run the algorithm fast and gives its good approximation.

Following minimal set of rules generated for fake and genuine review detection.

If(username &&review_sentiment&&review_category) then review = genuine.

If(username &&review_length&&review_sentiment) then Review=genuine

If(username &&review_sentiment&& rating) then Review= genuine

3.2 Decision tree

Decision tree algorithm takes input from the active learning labeled data. It takes each review in recursive manner and incorporates each parameter of active learning for arriving at conclusion. It recursively compares each parameter and results into true or false branch. Leaf node has node children and it indicates the decision of genuine or fake review.

Figure 1 shows each node makes decision based on following rules. Rules are applied in order to make proper decision -

If review has been given by genuine user then he/she must have proper username. Here we check if username is non-numeric. If it is numeric it means user is not valid.

Next we will check if customer has given only star rating. If it is so we consider it as fake one as genuine user definitely want to say something about the product he has bought. Else we check next condition.

Here we check if sentiment analysis of review matches with the review rating the customer has given. For ex if customer said product is excellent but rating given by him is less than 3, then we comes to conclusion that review given is fake or simply not genuine.

Next thing to check is making sure that review given must contain words or literals related to the category of product. Need of checking this is to make sure that review text is genuine not an advertising .

For applying decision tree one must design proper rules for making decision at each node to get accurate result. A decision tree will be built using the whole dataset taking into consideration all parameters.

3.3 Random Forest

Random forest algorithm is a supervised classification algorithm. The main concern of this algorithm is to create the forest with a number of trees. In general, the more trees in the forest ,the more robust the forest looks like. In the same way in the random forest classifier, higher the number of trees in the forest, higher is the accuracy. It can produce efficient results on big databases. We need not to remove any variable while handling thousands of input variables. It gives approximations of what factors should get more importance in the classification. It creates an inner unbiased estimate of the simplification error as the forest building growths. It has an effective method for estimating lost data and maintains accuracy when a large proportion of the data are missing.It can handle thousands of input variables without variable removal. It gives approximations of what variables are important in the classification. It has methods for balancing error in class people unbalanced data sets. Forests that get constructed can be used in future and hence must be saved. Relations between variables and their classification produces prototypes & also provide data about the relation. It finds neighborhoods between pairs of cases that can be used in clustering, locating deviations, or (by scaling) results into striking perception of the data. The abilities of the above can be extended to unlabeled data, leading to unsupervised clustering, data perceptions and deviation finding. It offers an experimental method for noticing variable interactions. Random forests does not over fit.

Algorithm of Random Forest

- Randomly select “K” features from total “m” features where k << m.
- Among the “K” features, calculate the node “d” using the best split point.
- Split the node into daughter nodes using the best split.
- Repeat the 1 to 3 steps until “l” number of nodes has been reached
- Build forest by iterating steps 1 to 4 for “n” number of times thereby creating “n” number of trees

Random forest prediction pseudocode:

To perform prediction using the trained random forest algorithm uses the below pseudocode.

- Takes the test features and use the rules of each randomly created decision tree to predict the oucome and stores the predicted outcome (target)
- Calculate the votes for each predicted target.
- Consider the high voted predicted target as the final prediction from the random forest algorithm.

Here out of 6 features, we tried to combine few features at a time that produces number of trees in a forest.

If Category =1 && rating=1 then tree1++.
If review_text comparison =1 && rating=1 then tree2++.
If repeated_review =1 then tree3++
If review_length =1 then tree4++
If only rating =1 then tree5++

Thus number of trees created for different features andwe get the count for genuine and fake review for each tree in a forest. And finally we find probability of combined features/factors.

3.4 Support Vector Machine

SVM is more useful in classification problems. This algorithm allows us to plot each data item as a point in n-dimensional space (where n is number of features used for classification) where the value of each feature indicate the value of a particular coordinate. Support vector machine is highly preferred by many because of its outstanding performance in producing accurate result with less computation power. A Support Vector Machine (SVM) classifier is judiciously used for separating hyperplane. In other words, given labeled training data *(supervised learning*), the algorithm produces an optimal hyperplane which categorizes each new incoming review into two classes. Hyperplane divides two dimensional space in two parts where in each class lay in either side.

Tuning parameters used for fake review detection in SVM are:

1) Kernel
2) Gamma

1) Kernel

The learning of the hyperplane in linear SVM is done by transforming the problem using some linear algebra. This is where the kernel plays role.

For linear kernel the equation for prediction for a new input using the dot product between the input (x) and each support vector (xi) is calculated as follows:

$F(x) = B(0) + sum(ai * (x, xi))$

Eq (1)

This is an equation that involves calculating the inner products of a new input vector (x) with all support vectors in training data. The coefficients B0 and ai (for each input) must be estimated from the training data by the learning algorithm.

The polynomial kernel can be written as

$K(x,xi) = 1 + sum(x * xi)$ ^d and

Eq(2)

exponential as

$K(x, xi) = exp(-gamma * sum((x—xi^2))$.

Eq (3)

Polynomial and exponential kernels calculates separation line in higher dimension. This is called kernel trick.

2) Gamma

The gamma parameter defines how far the influenceof a single training example reaches, with low values meaning ‘far’ and high values meaning ‘close’. In other words, with low gamma, points far away from plausible separation line are considered in calculation for the separation line. Whereas high gamma means the points close to plausible line are considered in calculation.

Here we divide our dataset of active learning parameters into two parts. First part contains 75% of dataset and second part 25% each of which is divided into training dataset and test dataset. This data is plotted on x and y axis. Hyperplane line is plotted with linear SVM. The datapoints of reviews that lies on or nearby hyperplane that is actual value are considered to be genuine else they counted as fake.

IV. APPLICATION SOFTWARE

In this study, reviews collected from customers are analyzed using different classifiers namely Roughset, Decision tree, Random forest and SVM. The dataset used in this research was downloaded from amazon online shopping site. Dataset contains total 1484 reviews. The Programming environment encompasses the java for building the application, jsp used for web based application and database stored in MySQL.

A. Preparation

- **Data Acquisition**- Acquire shopping database from online shopping site as amozon.com that contains reviews of customer. This data is unstructured and it is stored in xls format.
- **Data Preprocessing**- The database that is fetched into xls form is converted into structured database and stored in MySql format. In this step we clean the data comprising review comments. This step removes the stop words and performs stemming. Here we select the required data from review and store it into different parameters that has to be chosen for detecting fake review.
- **Active Learning**- This takes into account the parameters used for fake review detection. This step trains our model by taking into account parameters or factors. It labels the review based on all values hold by different identification factors or parameters. It includes tagging a label to the unprocessed data. In this step we make cluster head from the structured data. We create two clusters as “correct” and “incorrect”. Active learning is a special case of semi supervised machine learning which can interactively request the user to determine the class of some unknown data points to achieve the desired results.

Labeling the whole dataset manually is extremely time consuming and labor intensive. So, the algorithm actively queries the database entries for labeling the new data points.TF-IDF(“Term Frequency-Inverse Document Frequency”) of review is generated by multiplying the term frequency with the log of the ratio of the total number of reviews to the number of reviews in which the term appears.

TF-IDF can be determined by-

$TFIDF(n, d, D) = \sum TF(n,d) * IDF(n,D)$

Eq (4)

Here,

TF (n,d)= the number of times the term n appears in document d.

IDF (n,D)= log N

With, N = total number of documents in a collection = |D| | {d ∈D : n ∈d}| = number of documents where the term n appears.

After constructing the vectors using TF-IDF values, these sparse vectors are fed into the classifiers.

If username is non-numeric
value1=1
Else
value1=0
If review length >=5
value2=1
Else
value2=0
If review has rating plus review text
Value3=1
Else
Value3=0
If review Sentiment Vs Rating = Consistent
Value4=1
Else
Value4=0
If Review contains Category of product
Value5=1
Else
Value5=0
If review text is not similar to other reviews
Value6=1
Else
Value6=0

If all the six parameters are true then we label review as “correct” otherwise as “incorrect”.

Classifiers /Algorithms- Outcome of active learning is passed onto the classifiers to evaluate their accuracy-

- Rough Set Classifier
- Decision tree algorithm.
- Random forest algorithm
- Support Vector Machine

Result of all the classifiers fed for evaluating the accuracy of all the algorithms where it compares which algorithm shows promising results in terms of count of fake and genuine review and execution time.

B. Implementation of Model

The algorithm for the whole approach to detect review spam is given below-

- Acquire database and preprocess it to create Sparse Vector.
- Preprocessed data is passed onto classifiers.
- Classifiers classifies the reviews as fake or genuine
- Compare count of fake and genuine for each algorithm. Also compare execution time for each.
- accuracy = CLASSIFIER.accuracy
- EVALUATE classifier measuring accuracy

V. CONCLUSION & FUTURE WORK

This paper proposes an ensemble methodology for identifying Fake Review by renowned learning method (Active Learning) using real life data. We used several methods to analyze a dataset of Amazon product reviews. We worked on sentiment classification algorithms to apply a supervised learning on electronic products of amazon reviews. Our experimental approaches calculated the accuracy of Roughset Classifier, SVM (Support vector machine), Decision Tree and Random Forest i.e. sentiment classification algorithms. Additionally, we were able to classify how many given review are fake and genuine.

We have used various algorithms i.e. Roughset Classifier, SVM (Support vector machine), Decision Tree and Random Forest to identify fake and genuine reviews. We used three supervised learning algorithms to classify Sentiment of our dataset have been compared in this paper Roughset Classifier, SVM (Support vector machine), Decision Tree and Random Forest. Using the accuracy analysis for these three techniques, the measured results show that SVM algorithm outperforms other algorithms, and the most accurate for correctly classifying count of fake and genuine reviews. The method shows very promising results while conducting different experiments.

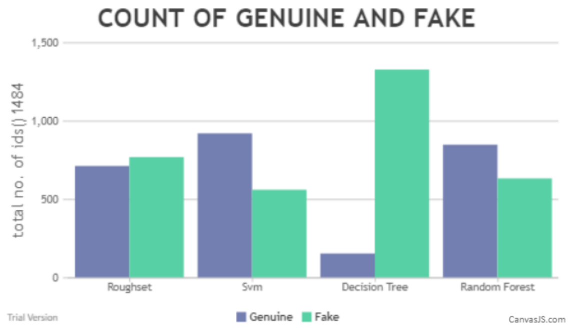

Fig 2: Comparative analysis of different algorithms

For future work, we would like to extend this study to use other datasets such as news spread by political parties and news channel during election days. Fake news detection takes into account different feature selection methods as news dataset consume different characteristics. Also, we can use different classification algorithm to find the rumors that goes viral in social media, which has high impact on society. we can use better sentiment classification algorithms to identify fake news using numerous tools such as Statistical Analysis System (SAS), Python and R studio.

Nowadays it is highly required to stop the fake news that get viral within very less time and it has very bad impact on the health of society. Rumors easily spread up because of social media. In country like India with huge population from different religions and being largest democracy, it is utmost important to maintain peaceful atmosphere. And hence identifying fake news that goes on viral is highly required. Social media constitutes different characteristics, so its quite confusing which factor helps to find fake news more promisingly and also without overfittting as some of the classifying algorithm tends to overfit in case of varying characteristics.

Just finding fake news will not be sufficient as current research is focusing highly on it. In future, finding fake news as early as possible is important because news and other things get viral so vastly and speedly and controlling them becomes a major issue. Thus finding fake news within a very less time will be a critical issue.

REFERENCES

[1] Parihar A., Bhagyanidhi, (2018) , “A Study on Sentiment Analysis ofProduct Reviews”, Soft-computing and Network Security(ICSNS) 2018 International Conference. pp. 1-5.

[2] Michael C., et al.,(2015), "Survey of review spam detection using machine learning techniques." Journal of Big Data 2.1, pp.9.

[3] Rajamohana S. P, Umamaheswari K., Dharani M., Vedackshya R.(2017), “Survey of review spam detection using machine learning techniques.”,978-1-5090-5778-8, pp.17 .

[4] Mevada D. L., Daxini V.,(2015), “An opinion spam analyzer for product Reviews using supervised machine Learning method.” pp.03.

[5] Adike R. G., Reddy V,.(2016), “Detection of Fake Review and Brand Spam Using Data Mining Technique.”, pp.02.

[6] Mukherjee A., VenkataramanV.,Bing Liu, Natalie Glance.,(2013),“Fake review detection: Classification and analysis of real and pseudo reviews”.

[7] Jiwei Li, MyleOtt, Cardie C., Hovy E.,(2014), “Towards a General Rule for Identifying Deceptive Opinion Spam”.

[8] Minqing H., Bing L.,(2004), “Mining and summarizing customer reviews”, Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, pp.168-177.

[9] Liu, Pan, et al. (2017),"Identifying Indicators of Fake Reviews Based on Spammer's Behavior Features." Software Quality, Reliability and Security Companion (QRS-C), IEEE International Conference, IEEE.

[10] Lim Ee-Peng, Nguyen Viet-An, Jindal Nitin, et al.(2010) “Detecting product review spammers

using rating behaviors”, Proceedings of the 19th ACM international conference on Information and knowledge management. New York: ACM Press, pp.939-948.

[11] [11] XieSihong, Guan W., Shuyang L., et al.(2012) “ Review spam detection via temporal pattern discovery” ,Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery And Data Mining. New York: ACM Press, pp.823-831.

[12] C.-H. Chu, C.-A.Wang, Y.-C.Chang, Y.-W.Wu, Y.-L.Hsieh, and W.-L. Hsu, (2016), “Sentiment analysis on chinese movie review with distributed keyword vector representation,” in Technologies and Applications of Artificial Intelligence (TAAI), Conferenceon.IEEE, pp.84–89.

[13] Bazan J.G., Nguyen H.S., Nguyen S.H., Synak P., Wróblewski J. (2000) Rough Set Algorithms in Classification Problem. In: Polkowski L., Tsumoto S., Lin T.Y. (eds) Rough Set Methods and Applications. Studies in Fuzziness and Soft Computing, vol 56. Physica, Heidelberg, 978-3-7908-1840-6, pp 49-88.