International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 07 | July 2020 www.irjet.net p-ISSN: 2395-0072

A Neural Network based Collaborative Approach for Spam Detection

Jay Shah[1], Dhruvil Jhaveri[2], Nitesh Hebbare[3], Jyothi Rao[4]

[1] *Student, Dept. of Computer Engineering, K.J Somaiya College of Engineering, Mumbai, India*
[2]*Student, Dept. of Computer Engineering, K.J Somaiya College of Engineering, Mumbai, India*
[3]*Student, Dept. of Computer Engineering, K.J Somaiya College of Engineering, Mumbai, India*
[4]*Professor, Dept. of Computer Engineering, K.J Somaiya College of Engineering, Mumbai, India*

Abstract - *While the digital world makes life easier people have become more and more dependent on technology applications. One of these is e-commerce websites where users purchase products or services at a certain facility and allow users to give appropriate feedback about that commodity in form of reviews. Such reviews provide valuable information sources for those products. Potential customers use them to get opinions from existing users before they decide to buy a product. Though use of such platforms also poses a challenge of authenticity and reliability of the reviews upon which many transactions are dependent. Unfortunately, this value of feedback often offers strong spam opportunities that contain false positive or malicious negative views .The proposed system uses text-based features such as noun-proportion and emotional analysis along with spam word count. By using these features we develop a neural network classification model to identify a review as spam or not spam. By evaluating our model on the yelp dataset of hotel reviews, and using the completely integrated neural networks , we found that experimental results obtained proves that our approach proposed outperforms current methods by attaining 94% accuracy .*

Key Words: **Spam review detection, Product review, Noun proportion, Emotional analysis, Spam dictionary**

1. INTRODUCTION

As the application of the internet is increasing, the usage of internet related applications is also growing. The quantity of user generated content in the form of review have significantly grown in recent past.[1] this is because the cost of internet connection has become cheaper at the same time the lives of people are becoming more and more busy which leads them to online shopping and online booking for services. The only relevant or reliable source for the people to rely upon while purchasing a product or booking for a service is the reviews recorded from the customers.[2]. At the same time there are many online competitors available in the market to provide the same service. This rivalry also leads to practises such as hiring third parties for writing biased reviews for satisfying the illicit means of the companies. These reviews can be biased in the favour of the same company or it can be against the rival company. These biased reviews can force users to buy the wrong product or it may also create wrong impressions about the product which will ultimately lead to the loss of buyer and the company.

Initially,[3] the three key kinds of spam opinions are: false opinions, brand-only criticism and non-opinions, along with other stumbling-up techniques. Many different researchers are trying to discover different things based on their research. Based on their work, several different researchers are seeking to discover the various capabilities of fake assessments and methods to clear up the issue successfully. Spam comments can be found by text mining and using Natural Language Processing (NLP). This includes methodologies such as calculation of the Word Frequency-Inverse Document Frequency (TF-IDF)[4], methods n- gram, etc. In such scenarios the need to develop a system which will protect the platform from such malpractices becomes important.

2. RELATED WORK

In recent years, plenty of studies have taken place for false review detection, i.e., exploring different machine learning approaches in many spam research. The relevant literature are as follows. In [15.] Many auto-generated spam comments are either filled in with some keywords in noun-phrase form without producing a full sentence, or links to keep the crawler going. These spam reviews are expected to have a higher concentration of the word noun. .Therefore, the idea here is that spam reviews have a higher noun concentration while non-spam reviews do not. In [18] the SVM model is used for decoding the sentiment tendency from word recurrence in a particular review. SVM models produced are compiled into a polarity index that shows the value of each sentence. Some studies, such as the study In[19], overlap semantic material to classify fake feedback with the aid of WordNet lexical databases

The research conducted in [11] is to detect duplicate reviews that are extremely close to duplicate. They calculate the text similarity score among the reviewers to detect this. To test the similarity of the message, KLD and JSD are used.
The work conducted in [2] aims at data collection and labelling is to artificially construct spam databases using simulated review spamming, making real reviews and making fake reviews. In [20], this captures the relations among all customers, feedback and items evaluated by the evaluators and expose the origin of spam.

So we have proposed a system that tries to classify the review written by the customer as authentic or not authentic. We used a Yelp data set to train our model. Many machine learning methodologies have been introduced to classify spam and non-spam, but very few of them are effective in classifying a review. We used spam word count, noun ratio, emotional score analysis features to train our model. We also referred to self-extensible spam dictionaries to increase spam terms in the dictionary and keep the dictionary current. We have used the technique of LSTM (long-term memory).Lastly we provide all the extracted features to Fully connected neural networks.

3. METHODOLOGY

The methodology proposed is based on five major features:-

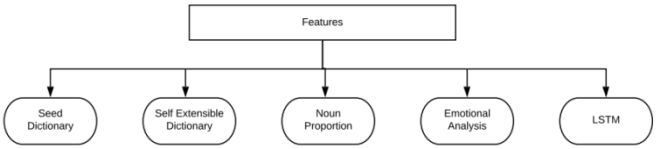

Fig - 1: Features proposition

- Seed Dictionary
- Self Extensible Dictionary
- Noun Proportion
- Emotional Analysis
- LSTM (Long Short Term Memory)

These five features form a base for generating a system for classification of review. Further these features also act as nodes in connected networks thereby increasing system efficiency of classification.

3.1 Feature Selection

Our approach is an in-depth study of spam comment identification. We integrate semantic analysis to create the self-extensible dictionary, which improves and extends with new cyber words. Using statistical analysis, we retrieve four text-based features, including repeated remarks, noun ratio, hyperlink number and emotional performance . These functions express differences between the ordinary and spam orders.

3.2 Dataset Selection

We use the first gold standard dataset created by Ott et al.[4][5]. The authors have created a negative opinion dataset for the detection of review spam in [4] and a positive opinion dataset in [5]. The dataset consists of 1600 reviews listing as, 400 disappointing negative reviews from Amazon Mechanical Turk, 400 true positive reviews from TripAdvisor, 400 false positive reviews from Amazon Mechanical Turk and 400 true negative reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor and Yelp.com. We split the dataset into the 8:2 ratio for training and testing, i.e. 80 % is used for training and 20 % remains for testing.

3.3 Feature Selection

For processing the data, it is important to clean the data in order to remove punctuation, marks, arrows, numerical characters from the reviews in the dataset and turn the outcome into a list so that each list can be further split into terms and phrases. Lemmatization is also carried out so that it resolves the word to it's dictionary form which further helps in emotional analysis. It is seen that regular expression is used for the processing of data. Further, building a seed dictionary containing unique terms in the analysis that will make it easier to extend the dictionary (See Fig. 3).

3.4 System Architecture

There are three major modules of this block diagram (see Fig. 2) :-

- Self Extensible Seed Dictionary
  - Gathering of Data
  - Cleansing of Data
  - Generation of seed dictionary from online sources and manually
  - Generation of self extensible dictionary using seed and review
- Text Based Features
  - Noun Proportion
  - Emotional Analysis
- Long short-term memory

Further these sections form the nodes of FULLY CONNECTED neural networks and act as a base for classification of review. Finally the review is classified as spam or not.

In this approach a seed dictionary is created initially and compared with reviews of the dataset to identify whether the review is spam or not, although classification is taking place but it is not accurate so in order to resolve the issue text based features are used. Hence all of these act as a node for a fully connected network.

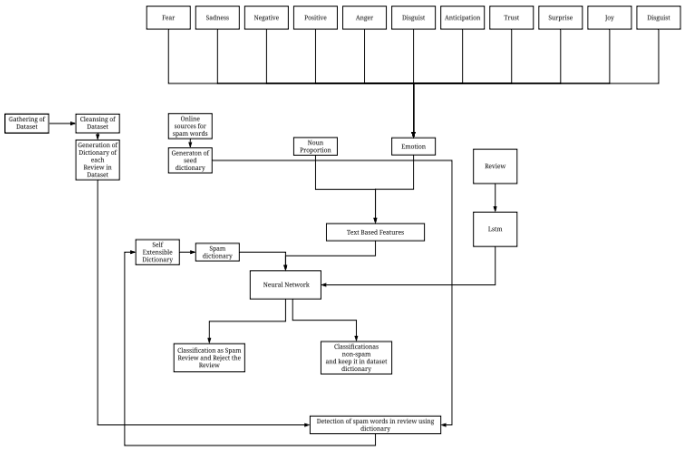

Fig - 2: System Architecture

3.5 Extensible Spam Dictionary

The dictionary ensures that as more the reviews are compared with the initial seed dictionary, this dictionary extends further and finds similar terms to that spam word in the review and adds it to the dictionary which is also a spam term, so that it can be used further down the to compare with other reviews and to reach maximum accuracy. (see Fig. 4)

In this word2vec model is used which consists of a thousand word vectors and the corpus is easily accessible online. It is used to introduce more terms to the dictionary as it incorporates the similarity of almost all terms identical to a single word. That is, happiness is important to be happy, cheerful etc. (see Fig - 4)
Algorithm to obtain the self extensible dictionary is as follows [6]: -

- Selecting the review
- Converting the review elements into a vector
- Finding the dissimilarity between the word in seed spam dictionary and in reviews
- Converting each word of the sentence into vector
- Sorting the dissimilarity weights of the words from seed words for the entire review
- Appending 15 most similar words
- If there are 3 or more than 3 words in the seed dictionary then add the word of candidate spam dictionary to the seed spam dictionary.
- Convert the added word into a vector.

3.6 Emotional Score

The main aim of the spammer is to post reviews that look similar to a real review. But the intentions behind spamming are expressed in the emotions. If the spammer tries to promote a product the hidden emotion would be of joy, trust, positive, etc. On the other hand if he tries to demote product emotions would be most likely of anger, disgust, and sadness, negative. The review of balanced emotions is less likely to be fake. We first find out the emotional diversity of the reviews by finding the score of each emotion for each review. This is done with the help of NRC Emotion Lexicon [17]. Different emotions like anger, disgust, joy, sadness, fear, etc are used as parameters. To calculate the score for a particular emotion [16].

Score (i,j) = Num(Words(j)) / Numword(i)

Where i is the review and j is the emotion. (i) is the total of words in review. (Words(j)) is the total words in the review that express the emotion j. This score is calculated for each emotion and for the entire (see Fig. 6). To calculate accuracy obtained by this method, the score becomes nodes in the model.

3.7 Noun Proportion

A summary consists of nouns, verbs, adjectives, conjunctions, etc., regarded as a bag of words, although studies have shown that a spam summary consists of a high percentage of nouns. There are essentially four forms of nouns:-

- NN noun singular
- NNS noun plural
- NNP proper noun singular
- NNPS proper noun plural

So in order to calculate the noun proportion of a each particular review in a dataset following steps are as follows:

- For Each Text review in dataset
  - Store each review in text
  - Inserting the text review into textBlob() function
  - Tagging each word's part of Speech
  - Keep a track of total number of words in the review
  - Keep a track of all the noun occurring in the review
  - Calculating the noun percentage
- Run test() function and Output

There are reviews which contain more than 40% noun proportion in a review which is most likely to be considered as spam but we still cannot classify the review as it can either be positive or negative (see Fig. 5) .

3.8 Bidirectional LSTM

Recurrent Neural Networks (RNN) are well-known models that have proven themselves in many NLP tasks. RNN performs the same job repetition for each element of the input sequence and output is calculated on the basis of previous calculations. LSTM networks have certain state

cells which act as long-term or short-term memories. The state of those cells modulates the output of the LSTM network. We use LSTM because we need neural network prediction to rely on the historical background of inputs, rather than just the very last input.

Instead of using traditional LSTM we are using a more powerful version of it i.e. Bidirectional LSTM. It will feed in the input sequence as usual for the first LSTM network. It will invert the input sequence and feed into the LSTM network for the second LSTM network. These two networks' output will be merged, and then passed on to the next layer. The sentences we pass through the model are not of the same size, so we do sequencing and padding to generate proper LSTM model efficiency. In the classification model, the output is further utilized as a node.

3.9 Classification Model

We have used a fully connected neural network in which neurons between two adjacent layers are fully connected in pairs. Here each output dimension depends upon each input dimension. We have used deeper networks because it can interpret complex functions more efficiently. The initial nodes in the neural network that we use are the scores from self extensible dictionary, emotional score, noun proportion and LSTM. We used 4 dense layers.

Dropout is a form of regularization. The dropout layer is used after every dense layer because it prevents the network from memorizing the training data. It also helps to boost the predicting ability of the model for the new data. The dropout is 25%. For each dense layer the activation feature ReLu is used. We've used sigmoid as an activation feature for the final layer. Binary cross entropy is used as a function of failure. Rmsprop is used as an optimizer. With this setting of all the hyper parameters we obtain the best precision (see Fig. 7).

4. RESULTS

4.1 Generation of Unique words

Words obtained are unique words which exist in the reviews of these can be useful for generation of initial seed dictionaries.

Fig - 3: Unique Words in the review

As we see in above Fig -, words generated are a list of all the unique words of a review in the entire dataset , Further these words can be used to create a seed dictionary for the initial stage of the system .

4.2 Self - extensible Dictionary

As we see in the original dictionary, there were 70 words but 255 more new words were added to the dictionary after expanding the dictionary. Furthermore, when comparing the initial seed dictionary and the extensible seed dictionary on the analysis for the classification, the initial seed dictionary is found to be less accurate.

```
Candidate Spam dictionary: ['stupidity', 'replied', 'collect']
iteration=1226
Candidate Spam dictionary: []
iteration=1227
Candidate Spam dictionary: []
```

Fig - 4: Addition of words in seed dictionary

Above Fig - is a snippet depicting addition of new words in the dictionary after a series of iterations on the dataset . In iteration 1226 few words are added but it does not happen for all the iterations.

Accuracy of classification of review which was 48.76 % using the initial dictionary is improved to 49.38 % when the dictionary is extended and loaded to machine for classification.

4.3 Noun Proportion

Spam review contains a high score of nouns in it so in order for classification of review generating noun score of reviews will elevate the efficiency to classify the review

```
1=>positive=>truthful=>40.0
60=>positive=>truthful=>42.857142857142854
271=>positive=>truthful=>40.42553191489362
277=>positive=>truthful=>41.935483870967744
333=>positive=>truthful=>40.0
398=>positive=>truthful=>41.666666666666664
666=>positive=>deceptive=>83.13253012048193
771=>positive=>deceptive=>43.70860927152318
990=>negative=>truthful=>40.0
```

Fig - 5: Noun Proportion of review

In above Fig - percentage of noun is calculated resulting in which will give an estimation for classification of reviews. As we can see review 666 contains 81% of nouns in it which means that review is spam and so for other reviews in the dataset as follows .

4.4 Emotional Analysis

Conducting sentiment analysis on a review will give a broader aspect about the polarity of review i.e positive or negative. Also it tells about different emotions that could be detected from the review.

	anger	fear	anticipation	...	disgust	positive	negative
	0.009524	0.009524	0.019048	...	0.0	0.057143	0.028571

Fig - 6: Sentiment Analysis of review

In the above Fig - an emotional score is calculated for each emotion. This results in the summation which will give a positive and negative sentiment score which will help in classifying the review as spam or not . Hence here review is positive as it scores 0.057 in positive and scores 0.01 in anticipation and very low in anger and fear i.e 0.009.

4.5 System Performance

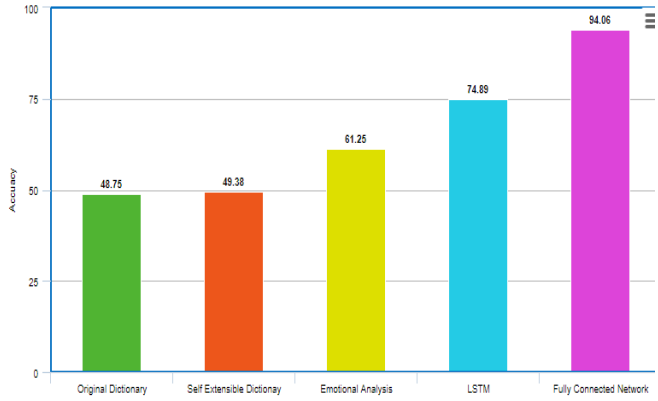

Fig - 7: Final Accuracy

In above Fig - we can see that when features (i.e original dictionary , Extensible dictionary , Emotional Score , LSTM) are trained individually on the system there is specific accuracy attained (i.e. 48.75 , 49.38 , 61.25 , 74.98 respectively) and at last when there are all combined (i.e Fully Connected Network) accuracy is boost up to 94.06%

5. CONCLUSION AND FUTURE SCOPE

As we know users can spam unnecessary reviews on a particular product/website to either degrade the performance or to spoil its image by which other users will not get attracted towards the product/website. In this paper, we introduce textual analysis to create a self-extended dictionary that updates and expands with new spam words. We extract text-based features like noun ratio, and emotional ranking, through quantitative analysis. These characteristics describe the distinctions between spam and non-spam reviews. Based on features above, we detect spam reviews. As above methods are not as efficient we have also used Neural Network models like LSTM and fully connected models to classify the spam reviews which has given us a greater system efficiency for classification of reviews. Although there is one limitation to this approach which is this system trains on labeled datasets, we conduct an in-depth analysis of detection of spam reviews which will help products/websites to flourish. While our approach's experimental results are acceptable for spam reviews detection, useful suggestions are available to improve the outcome.

There is a need to have more research on the detection of spam for multilingual reviews as there are very few systems for detection of multilingual reviews. To improve the accuracy of the algorithms there is a need for more attributes, such as the IP address of the spammer, email address for the review website, and location where the reviewer signed in to write the review.

REFERENCES

[1] 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 , ACL

[2] Dixit S, Agrawal AJ (2013) Survey on review spam detection. Int J Comput Commun Technol ISSN (PRINT) 4:0975–7449

[3] https://www.competitionbureau.gc.ca/eic/site/cb-bc.nsf/eng/03782.html - Not all reviews are from legitimate consumers

[4] https://consumerist.com/2010/04/14/how-you-spot-fake-online-reviews/ - Ways You Can Spot Fake Online Reviews

[5] https://tripadvisor.mediaroom.com/us - Review Database

[6] Spam Comments Detection with Self-Extensible Dictionary and Text-Based Features (2017) Qiang Zhang, Chenwei Liu, Shangru Zhong, Kai Lei* Institute of Big Data Technologies Shenzhen Key Lab for Cloud Computing Technology & Applications School of Electronics and Computer Engineering(SECE) Peking University, SHENZHEN 518055 P.R.CHINA

[7] Fangzhao Wu, Jinyun Shu, Yongfeng Huang, and Zhigang Yuan. Co-detecting social spammers and spam messages in microblogging via exploiting social contexts. In Neurocomputing, volume 201, 2016.

[8] Tomas Mikolov, Kai Chen, et al. word2vec, 2014. - Word to Vector Model

[9] Chenwei Liu, Jiawei Wang, and Kai Lei. Detecting spam comments posted in micro-blogs using the self-extensible spam dictionary. In 2016 IEEE International Conference on Communications (ICC), pages 1–7. IEEE, 2016.

[10] 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.

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

[12] B. Bigi, "Using Kullback-Leibler distance for text categorization," in Proceeding ECIR 03 Proc. 25th Eur. Conf. IR Res., 2013, pp. 305– 319.

[13] Rohit Giyanani, Mukti Desai, "Spam Detection using Natural Language Processing", IOSR Journal of Computer Engineering (IOSR-JCE), Volume 16, Issue 5, Ver. IV (Sep – Oct. 2014), PP 116-119.
[14] Settles, Burr. "Active learning literature survey." University of Wisconsin, Madison , 2010, 52.55-66 : 11.
[15] Archana Bhattarai, Vasile Rus, and Dipankar Dasgupta "Characterizing Comment Spam in the Blogosphere through Content Analysis ".15 May 2009
[16] Crowdsourcing a Word-Emotion Association Lexicon, Saif Mohammad and Peter Turney, Computational Intelligence, 29 (3), 436-465, 2013
[17] Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon, Saif Mohammad and Peter Turney, In Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, June 2010, LA, California.
[18] Yu, Boya; Zhou, Jiaxu; Zhang, Yi; Cao, Yunong, "Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews," ARXIV, 2017.
[19] R. Y. K. Lau, S. Y. Liao, R. C. W. Kwok, K. Xu, Y. Xia, and Y. Li, "Text mining and probabilistic language modeling for online review spam detecting," ACM Transactions on Management Information Systems, vol. 2, Dec. 2011, pp. 1-30, doi:10.1145/2070710.2070716
[20] Jindal Nitin, Liu Bing, Lim Ee-peng, et al. " Finding unusual review patterns using unexpected rules", Proceedings of the 19th ACM International Conference on Information and Knowledge Management. New York: ACM Press -2010 ; pp 1549-1552

BIOGRAPHIES

Mr. Jay Shah is a senior undergraduate student pursuing Bachelor of Technology focused in Computer Engineering at K.J. Somaiya College of Engineering, Mumbai. His main areas of interest are Software Development, Machine Learning, Artificial Intelligence and Deep Learning.

Mr. Dhruvil Jhaveri is a senior undergraduate student pursuing Bachelor of Technology focused in Computer Engineering at K.J. Somaiya College of Engineering, Mumbai. His main areas of interest are Security, Software Development, Machine Learning, Deep learning applications like NLP.

Mr. Nitesh Hebbare is a senior undergraduate student pursuing Bachelor of Technology focused in Computer Engineering at K.J. Somaiya College of Engineering, Mumbai. His main areas of interest are Machine Learning and Deep learning applications.

Prof Jyothi M. Rao is working as Associate Professor in Department of Computer Engineering at K.J.Somaiya College of Engineering., Vidyavihar. Her main areas of expertise are Data Mining and Data Analytics. She is among the top 1% of NPTEL Big Data Online Certification.