International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-5, June 2020

Opinion Based Ranking System using Machine Learning

Manisha Chaudhari, Niranjan K. Prajapati, Rama M. Maliya

Abstract: ***The rapid increase in the online services in the recent years. Everyone sending feedback/review after used particular services. For unstructured data it released active countless opportunity ties and challenges for data mining research. This paper is especially for reviews of hotels which are given by various hotels visitors. Reviews are posted as a comment only. It is difficult to identify the positive & negative review. We used dataset of different hotels and perform sentiment analysis process. For classification word to Vec Algorithm is being used. For the positive and negative review calculation r2 & f1 scoring functions are very useful.***

Index Terms: ***Sentiment Analysis, Machine Learning, Deep Leaning, opinion mining, CNN, RNN, SVM, NB.***

I. INTRODUCTION

Now a days almost people are giving reviews after used hotel/other services. This is a big problem for us when we are trying to find good review, because of all reviews are together. For the classify both review we must have to use particular technique for the same. And Text classification offers good framework for getting familiar with textual data processing without lacking interest. Sentiment analysis aims to estimate the sentiment polarity of a body of the text based on its content. The sentiment separation of text can be well-defined as a value that says whether the stated opinion is positive (polarity=1) and negative (polarity=0). We proposed the system specify the posted hotel is good, bad based on opinions with rank.System will use dataset and will match the review with the sentiment keywords in dataset and rank the review accordingly. System will rate the hotel based on the rank of review.Different ways used to do Sentiment Analysis : 1. Rule Based Approach 2. Automatic Approach 3. Hybrid Approach

Automatic Sentiment Analysis methods are differing to rule-based systems even it doesn't depend on manually created rules, but it depends on machine learning techniques. Train a model of classification problems is main part of sentiment analysis where a fed with a text and the corresponding category is return by a classifier, e.g. positive, negative.

Classification Algorithms

1. *Naïve Bayes*
2. *Linear Regression*
3. *Support Vector Machines*
4. *Deep Learning*

Deep Learning trying to set artificial neural network process data to human brain by using different set of algorithms. For the sentiment analysis we need to train a Data Model by using deep learning techniques. This model process involved six different steps. Steps are Get Data, Generate Embedding, Model Architecture, Model Parameters, Train and test the Model and last step is Run the Model. Traditional machine learning algorithms require few training data compare to deep learning algorithms. traditional machine learning algorithms Support Vector Machine (SVM) and Naïve Bayes (NB) are useful. SVM and NB gives a accurate result for text classification. Even it doesn't requires much more data.

Advantages:

- It gives high accuracy for text classification with less data.
- Deep learning classifiers continue to get better the more data

As we know traditional algorithms gives accuracy for less data only. Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are deep learning architectures useful for text classification. These architectures are important and very useful for getting accurate result with large amount of data. Even we need to train a model for large amount of a data, because of the hotel ranking process include so much reviews.

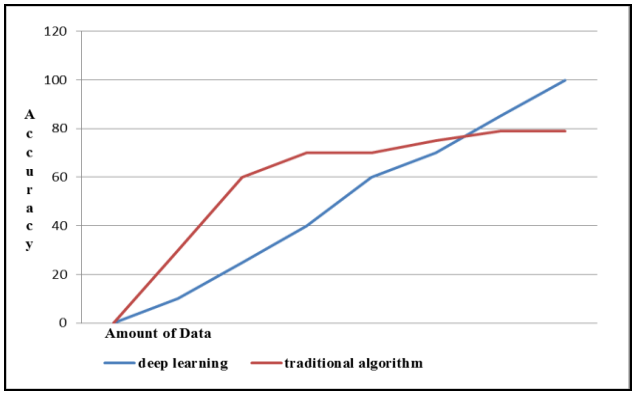

Figure 1.1: deep learning technique performance

In the above graph we can see clearly deep earning technique performance is very accurate with the large amount of data compare to traditional algorithm. Traditional algorithm also gives use accurate result but only with less data. So hotel ranking process only deep learning technique is useful for this research work.

Revised Manuscript Received on May 25, 2020.
* Correspondence Author
Mrs. Manisha M. Chaudhari*, Lecturer, Department of Computer Engineering, Government Engineering College, Gandhinagar, India.
Prof. Niranjan Prajapati, Assistant Professor, Department of Computer Engineering, Government Engineering College, Gandhinagar, India.
Ms. Rama M. Maliya, Assistant Professor, Department of Computer Engineering, Government Engineering College, Gandhinagar, India.

II. REVIEW OF LITERATURE

Comparative Study On Online Review Of Different Types Of Hotel [3]:

This paper is all about customer's review for different hotel. In recent years almost customer's uses ecommerce platforms frequently. So many researchers also working on the same topic. Almost researcher only consider reviews instead of type of reviews. As we know review is given in the textual format even posted as a comments which is an unstructured data format. This author has tried to extract unstructured data & convert into structured format So, it will be easy for the customers to use any online services. For the analysis of unstructured data, Natural Language Processing technique is very useful. And term frequency–inverse document frequency (TF-IDF) is used for customers level for comments.

It will generate a difference on data characteristics of different reviews of different hotel types.

A Survey on Fake Review Detection Using Machine Learning [2]:

This paper is about fake review detection. We know very well now a days online marketing & online shopping in trends due people don't have time to go shop & buy particular product. When we want to buy any product at that time we are just reading reviews of given by other customers for the same product & we take decision weather this product is suitable for us or not. After delivery we are not getting satisfied product as per reviews given on the website. Reason behind this case is just reviews are fake. So many companies are hiring employee for the post positive reviews on their product for their competition in the market with other competitors. Solution of this problem is only we should detect such kind of reviews. Now question is how to detect such fake reviews?? We can use data mining techniques for this detection & remove the fake reviews. Data Mining techniques like.. Supervised, Unsupervised and semi supervised technique features are useful.

Disaggregate Hotel Evaluation by using Diverse Aspects from User Reviews [1]:

In this paper author had research about hotel reviews. For the discovering coherent hotel aspects unsupervised method is useful. This model integrates two techniques 1. Modelling and 2. To automatically discover coherent hotel aspects word embedding, with the frequent noun-adjective co-occurrence statistics are needed. And for the hotel ranking process Supervised method is useful. Any predefined words are not being used by this method. and user attention with integrated techniques modeling and word embedding directly using for notices coherent level aspects. This method performing evaluation task by gathering different hotel reviews from lots of travel websites. Resultant baseline improved performance is increased and it is up to 90%. The result obtained are very favorable and show that the system is very simple, scalable.

III. PROPOSED WORK

Opinion Based Hotel Ranking Using Machine Learning Approach, for the hotel ranking we will use Sentiment Analysis approach. During the analysis Reviews will be divided into two parts positive & negative by using Deep Learning classification algorithm.

This process will complete in different following stages:

1. Import Dataset
2. Clean Text
3. Extract Truth Value
4. Shuffling Data
5. Importing pre-processed data in Pandas Data-Frames.
6. Create a vocabulary
7. Extract Embeddings matrix
8. Preprocess training text
9. Train the model
10. Test the model
11. Confusion matrix
12. Classification report
13. Calculate positive & negative scores
14. Extracting sentiment classes
15. Analysis of best to worst review

The process during above steps is given below in details :

The multiplayer perceptron will be built with **tensorflow keras** libraries. Useful function from **sklearn** will be used to evaluate the performances. Google's Word2Vec skipgram model will be extracted from **gensim libraries**. **Numpy** will be used to create the embeddings matrix of weigths out of the skipgram model.

To train the multilayer perceptron model we need to extract reviews text and assign them a truth value.

"Bad Hotel" : 0
"Good Hotel" : 1

Negative and positive reviews are already divided in the given dataset, so we just need to read data and create new columns with truth values.

Two dataset will be created:

1. sentiment_task_reviews, containing all reviews with truth
2. reviews_text containing all reviews (to train embeddings)

The **Word2Vec skip-gram** model is a simple Neural Network that performs a *fake task* of predict the nearest words given another one that appears in a sentence. The goal of this fake task is to train the network so that we can then extract the weights matrix in which each row will be the vectorial representation of a word.
We will train the network 128 neurons in the hidden layer, so the words will have 128 dimensional representations.

A **Keras Tokenizer** will extract the vocabulary of all words that appear in the 515000 reviews.

The word2vec model is trained we can combine its

vectorial representation of words to create and embeddings matrix:

vocabularySize x embeddingDim

After the preproccing we will get 90% accuracy easily. This final visualization show us that while the big part of positive scores seems to tend to 1, negative once seem to be more distributed in the [0, 0.5] range. This could mean that a big part of negative reviews are not so negative in the end. Maybe some guests who leave a positive review, leave also an "advice" as negative one.

Extracting sentiment classes

We're going to divide reviews in four classes:

1. **best** : fnl_score >= 0.7
2. **good** : fnl_score < 0.7 AND fnl_score >= 0.5
3. **bad** : fnl_score < 0.5 AND fnl_score >= 0.3
4. **worst** : fnl_score < 0.3

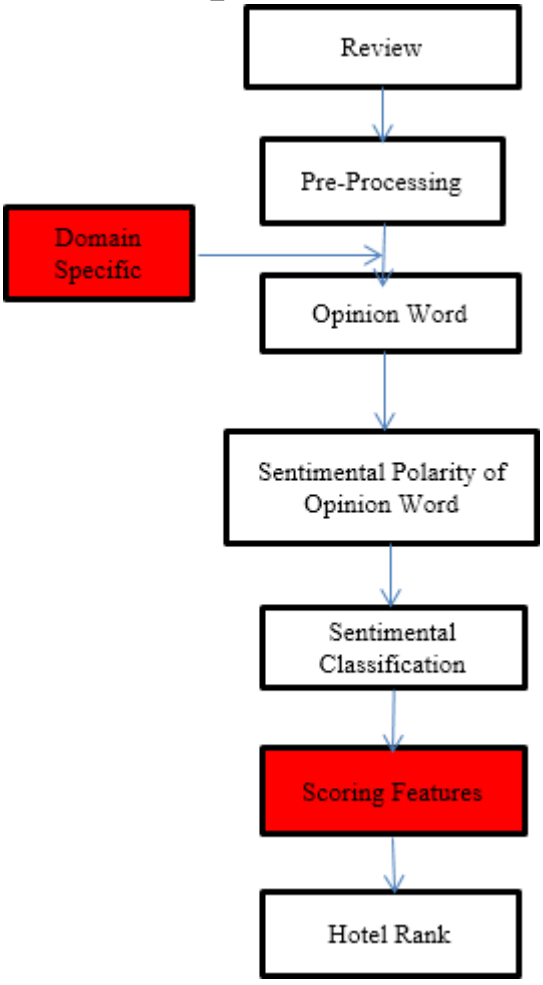

Figure 1.2 : system flow

Word to Vec Algorithm

Step 1: Data Preparation

In this step first define corpus, cleaning data, normalise and tokenise words performed.

Step-2: Hyper parameters

Learning rate, epochs, window size, embedding size for different parameters.

Step-3: Generate Training Data

Construct vocabulary and one hot encrypting for words, construct dictionaries that draw ID to word and word to ID.

Step-4: Model Training

Pass encrypted words through forward pass and calculate rate of errors and adjust weights using back propagation and calculate loss.

Step-5: Inference

Get word vector and find other words which are similar to them.

Step-6: Further improvements

For the Rank calculation uses of scoring functions like f1, r2 etc..

IV. RESULT

- using r2 score function

```
report = r2_score(y_test, predictions,multioutput='variance_weighted')
print(report)
report = r2_score(y_test, predictions,multioutput='uniform_average')
print(report)
report = r2_score(y_test, predictions,multioutput='raw_values')
print(report)
```

```
0.7265832128686986
0.7265832128686986
[0.72658321]
```

[Figure 1.3: r2 score function result]

- using f1 score function

```
report = classification_report(y_test, predictions, target_names=['0','1'])
print(report)
```

```
              precision    recall  f1-score   support

           0       0.91      0.94      0.93      2501
           1       0.95      0.93      0.94      3055

   micro avg       0.93      0.93      0.93      5556
   macro avg       0.93      0.93      0.93      5556
weighted avg       0.93      0.93      0.93      5556
```

[Figure 1.4: f1 score function result]

As we can see in the above figures f1 function gives us accurate result compare to f1 function.

V. MAJOR CHALLENGES IN HOTEL RANKING

The challenges which are involved in different opinion/review are below.

1. Sometimes review is given with only star rating at this time it is difficult to identify whether it is positive or negative[2].
2. When review is fake it is difficult to find out fake reviews.

VI. CONCLUSION

Opinion based Ranking System can improve the quality of hotels which are actually not good.

Good & Bad both reviews system will predict so customer can book hotel on the based of given +ve or –ve reviews. The system is likely to define all the words from the dataset using sentiment analysis & system detect it automatically and at the end of analysis system will generate csv file of positive & negative reviews.

REFERENCES

1. Bidur Devkota, Chenyi Zhuang, Kyoung-Sook Kim and Hiroyuki Miyazaki, Disaggregate Hotel Evaluation by using Diverse Aspects from publication (User Reviews, 2019).
2. Nidhi A. Patel and Prof. Rakesh Patel, A Survey on Fake Review Detection using Machine Learning Techniques, 2018, ICCCA, by IEEE, Greater Noida, India,
3. Tianjiao Niu, Yusi Ding and Jianzheng Yang , Comparative Study on Online Review of Different Types of Hotel, 2018, DSA by IEEE, Dalian-China.
4. Madan Lal Yadav and Basav Roychoudhury , Effect of trip mode on opinion about hotel aspects: A social media analysis Approach, 2019, IJHM by Elsevier publication, Shillong, Meghalaya India.
5. M.R. Martinez-Torres and S.L. Toral, A machine learning approach for the identification of the deceptive reviews in the hospitality sector using unique attributes and sentiment orientation, 2019, TM by Elsevier Publication, Spain.
6. Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yaser Jararweh and Omar Qawasmeh, Enhancing Aspect-Based Sentiment Analysis of Arabic Hotels' reviews using morphological, syntactic and semantic features, 2016, IPM by Research Gate,.
7. Kudakwas he Zvarevashe and Oludayo O Olugbara, A Framework for Sentiment Analysis with Opinion Mining of Hotel Reviews , 2018, ICTAS by IEEE, Durban, South Africa.
8. Jian-Qiang Wang, Xu Zhang and Hong-Yu Zhang, Hotel recommendation approach based on the online consumer reviews using interval neutrosophic linguistic numbers, 2018, I&FS by IOS press & content library, Changsha 410083, China.

AUTHORS PROFILE

Mrs. Manisha M. Chaudhari is working as a Lecturer & right now doing research as a part of Master's Degree. She has done Bachelor Degree in the field of Computer Engineering & having more than 9 years of teaching experience, she is doing research in machine learning area & lifetime member of IFERP.

Prof. Niranjan Prajapati is working as an Assistant Professor at Government Engineering College Gandhinagar & guiding Bachelor's Master's Students for their project. He is professional member of ISTE. He is also having 7+ years bright experience in the teaching field.

Ms. Rama M. Maliya is an Assistant Professor at SAL Education having Master's Degree in Computer Engineering She is having 6+ years of teaching experience & also working as a freelancer in IT sector like, web development, social media handling etc. She is also lifetime member of IFERP. She has done her research work in VANET during master's degree & currently doing research in the DATA Mining field.