International Research Journal of Computer Science (IRJCS)
Issue 04, Volume 07 (April 2020)

ISSN: 2393-9842
www.irjcs.com

CUSTOMER REVIEW ANALYSIS - MULTI-LABEL CLASSIFICATION AND SENTIMENT ANALYSIS

Shreyas Renga
Computer Science & Anna University, INDIA
shreyasrenga99@gmail.com
Abishek Ganapathy
Computer Science & Anna University, INDIA
abishekganapathy15592@gmail.com
T. Hasith Ram Varma
Computer Science & Anna University, INDIA
hasithram@gmail.com

Manuscript History
Number: **IRJCS/RS/Vol.07/Issue04/APCS10082**
Received: 29, March 2020
Final Correction: 08, April 2020
Final Accepted: 19, April 2020
Published: **April 2020**
Citation: Shreyas, Abishek & Hasith, T. (2020). Customer Review Analysis - Multi-Label Classification and Sentiment Analysis. International Research Journal of Computer Science (IRJCS), Volume VII, 28-32.
doi://10.26562/IRJCS.2020.APCS10080
Editor: Dr.A.Arul L.S, Chief Editor, IRJCS, AM Publications, India

Abstract: In this paper we focus on the hotel sectors and help them process these huge chunks of data in the form of customer reviews and help them derive useful information. The data pre-processing involves the scrapping of reviews from different sites and storing them and also check the correctness of the regular expression of the reviews. Our modelling employed includes three machine learning algorithms namely Naive Bayes, Support vector machine (svm) and Logistic regression. These three models improve the accuracy of the model as well as its robustness. The main idea of using these models are that the reviews are labelled so that the hotel management need not waste loads of time reading all the reviews. Instead the important reviews can be arranged based on their polarity and the important topic discussed in the review can be highlighted. So that it is easy for the management to analyse both the positive as well as the negative reviews. Sentiment polarity is incorporated to arrange the reviews based on the sentiment the review establishes. This paper helps the world to properly analyse the feedbacks and the reviews given by the customers.

Keywords: analysis; review; customer; multi label; Naive Bayes; Support vector machine; Logistic regression; Sentiment polarity; Multi-label classification; POS tagging;Based ; Calibration; Computer; Study; Architecture;

I. INTRODUCTION

Nowadays people are in the urge of improving themselves. People strive to be a better person, do better in business and the chores they do. They improve themselves by asking feedback and reviews from others. But the problem arises when the numbers of feedback or reviews are more than the threshold of the person or the organisation. For handling these feedbacks, we have developed a customer review analysis system employing machine learning and natural language processing. The main objective of this paper is to create a customer review analysis system where we can help the concerned person to identify reviews of at most importance. The system we design will be more user friendly as they are intimidated only on the reviews of highest priority. The reviews are scrapped from all the review portals making it more efficient. The hotel management will benefit from the analysis that is provided because they can find the areas where they can improve. Identification of key topics in reviews.

This is a Multi-label classification problem, and each review could contain more than one topic[4]. Topics will be restricted to cleanliness, room service and location. Getting the sentiment polarity for each topic will be added value[1].

II. SYSTEM ARCHITECTURE

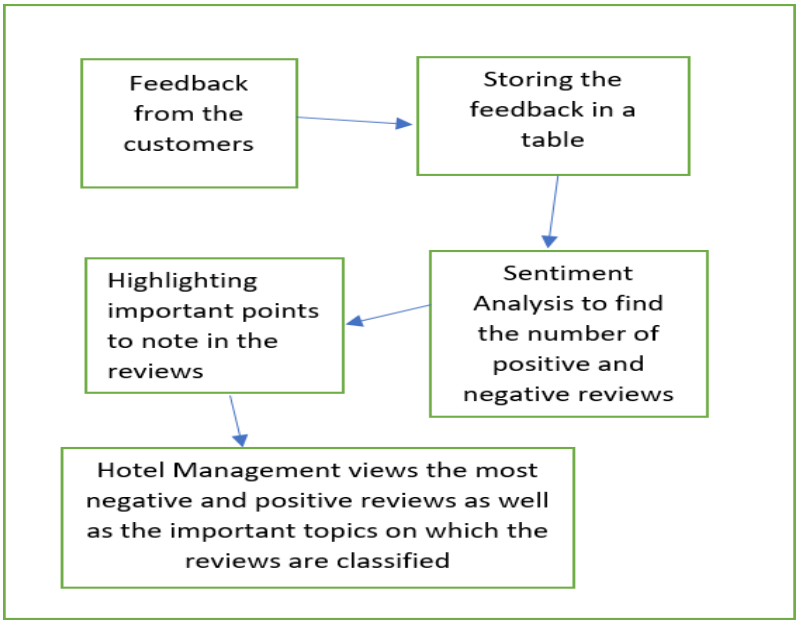

Fig.1 Proposed system block diagram

III. SORTING THE REVIEWS BASED ON SENTIMENT POLARITY

The sentiment of the reviews are calculated using the sentimentintensityanalyser. The sentiments of the reviews are classified into positive, negative and neutral. The positive reviews are the happy customers; the negative reviews denote the frustrated customers and neutral customers are in between these two sentiments. The top positive reviews and negative reviews are taken and are sorted in decreasing polarity, so that the hotel management can be indicated on both the most positive and negative values and highlighted using POS tagging [1].

	review_content	neg	pos	neu	compound
0	I visited Rainforest resort with ny family on ...	0.000	0.177	0.823	0.9778
1	Amazing staff, rooms, view and food. Could not...	0.047	0.305	0.647	0.9408
2	This resort has a multitude of restaurants, ba...	0.000	0.262	0.738	0.9946
3	The island is wonderful, the facilities in exc...	0.000	0.266	0.734	0.8883
4	The best place to relax!great restaurants,uniq...	0.000	0.337	0.663	0.9410

Fig.2 Customer Reviews with their respective sentiment polarity

IV. MULTI-LABEL CLASSIFICATION

There is always a confusion regarding the two terms: multilabel classification and multiclass classification. Multiclass classification refers to a classification task that generally has more than two classes where each sample that is classified can only have one class. For example: A fruit can be classified as either an orange or an apple but not both. Whereas Multilabel classification refers to the kind of classification where each sample is assigned multiple target labels[4]. For example: A text in a particular document can talk about the politics in sports where players are payed more money if they play very well in a match. Here the text can be assigned labels of Sports, Finance and Politics simultaneously. So, this customer review analysis system uses multilabel classification as each review given by the user can have multiple labels assigned to it.

This paper mainly considers three labels: Location, Cleanliness and Room Service. We assume that we have three bags corresponding to the three labels with sample words corresponding to the labels inside them that would help us in classification. This Multi-Label Classification can help the Hotel Management easily identify the topic the review lies on.

text	sentiment	class
ideal place	positive	location
ideal place to stay	positive	location
easy accessibility	positive	comfort & facilities
clean and nice room	positive	cleanliness
nice room	positive	comfort & facilities
unforgettable stay	positive	comfort & facilities
Lovely resort	positive	comfort & facilities
good staff	positive	staff
Clean resort	positive	cleanliness
friendly staff	positive	staff
great location	positive	location

Fig. 4 Class column denotes the Labels of the reviews

V. MODELS USED FOR TRAINING THE CUSTOMER REVIEWS

Model 1: Support Vector Machines (SVM)

The aim of Support Vector classification is to devise a computationally efficient way of learning 'good' separating hyperplanes in a high dimensional feature space, where by 'good' hyperplanes we will understand ones optimising the generalisation bounds, and by 'computationally efficient' we will mean algorithms able to deal with sample sizes of the order of 100,000 instances[2]. Simply speaking SVM tries to find a hypothesis for which we can guarantee lowest true error possible. We consider SVM for multilabel classification because as Joachims[3] suggests SVM uses overfitting protection and thus can handle large feature spaces.

Model 2: Naïve Bayes

Naive Bayes is a family of probabilistic algorithms that take advantage of probability theory and Bayes' Theorem to predict the tag of a text. They are probabilistic, which means that they calculate the probability of each tag for a given text, and then output the tag with the highest one. The "Naïve" term here refers to the assumption that every word in the sentence is independent of one another. This means that we are looking at individual words rather than individual sentences. So here comes the question of why to use Naïve Bayes for multilabel classification? The best solution would be that Naïve Bayes has a bias-variance trade-off. The text type data's are generally noisy and are of high dimensions. By assuming independence of words we're saying that covariance matrix of our model only has non-zero entries on the diagonal. This introduced bias may sufficiently reduce variance that you get better predictions.

From [4], we can infer that for short snippets, Multinomial NB does better than SVM and for longer texts, SVM does better than NB using 9 datasets (4 short texts, 5 long texts) with 2 classification tasks (6 sentiment classification tasks, 3 topic classification tasks).

VI. PSEUDOCODE FOR THE MODEL

Algorithm for assigning labels for each customer review

1. Start
2. locationWords = []
3. cleanlinessWords=[]
4. roomServiceWords=[]
5. firstLabel = ""
6. for every customer review:
    a. For each word in a review:
    b. Compare with the contents of each bag and find its label
    c. Add the word to its appropriate list
    d. If the word is the first word to be labelled then:

e. firstLabel = word's label
f. If len(locationWords)>len(cleanlinessWords):
g. If len(locationWords)>len(roomServiceWords):
h. Assign the entire review as location review
i. If len(cleanlinessWords)>len(locationWords):
j. If len(cleanlinessWords)>len(roomServiceWords):
k. Assign the entire review as Cleanliness review
l. If len(roomServiceWords)>len(locationWords):
m. If len(roomServiceWords)>len(cleanlinessWords):
n. Assign the entire review as Room Service review
o. If there exist multiple labels with same maximum list length:
p. The review is classified as that label which has been assigned first in the sentence among the tie-breaking labels

VII. PARTS OF SPEECH TAGGING

The main reason for using Parts Of Speech tagging or basically known as POS tagging, is that most of the models built without the help of POS tagging in the context of Natural Language processing do not have relationship between words, but in our case relationship between words has an great priority, and thus POS tagging plays a major role in Text Based Works [5]. Grammatical annotation can be useful in situations where you want to distinguish the grammatical functions of particular word forms or identify all the words performing a particular grammatical function [6].

For example, suppose we build a sentiment analyser based on only Bag of Words.
Such a model will not be able to capture the difference between "I like you", where "like" is a verb with a positive sentiment, and "I am like you", where "like" is a preposition with a neutral sentiment.

So this leaves us with a question — how do we improve on this Bag of Words technique?

1. Lexical Based Methods
2. Rule-Based Methods
3. Probabilistic Methods
4. Deep Learning Methods

In our case, the best one which we felt and used was the Rule-Based Methods, this assigns POS tags based on rules, which we have specified by adding some patterns for each rule, For example, we can have a rule that says, words ending with "ed" or "ing" must be assigned to a adjective.[6]. Rule-Based Techniques can be used along with Lexical Based approaches to allow POS Tagging of words that are not present in the training part but are there in the testing data. And thus formed patterns are added to the entity ruler, and we get the results, and there is also a chance that a tag assigned for the word at one place, might differ from other place, and that is done by the following Algorithm,[7]

Change tag 1 to tag 2 when:

1. The preceding (following) word is tagged m.
2. The word two before (after) is tagged m.
3. One of the two preceding (following) words is tagged M.
4. One of the three preceding (following) words is tagged m.
5. The preceding word is tagged z and the following word is tagged k.
6. The preceding (following) word is tagged m and the word two before (after) is tagged k.
7. The current word is (is not) capitalized.
8. The previous word is (is not) capitalized. And thus this algorithm helps us to change the tag, from 1 to 2 depending upon the sentence we have.

Fig. 5 A sample line graph using which shows the flow of POS tagging

Hence the total flow of the POS Tagging continues in the way showed above

The way for showing the pos tagged part in highlighted colours, is by using Displacy,
1. spaCy a inbuilt module for Natural Language Processing comes with a built-in dependency visualizer named Displacy that lets you check your model's predictions in your browser.
2. You can pass in one or more Doc objects and start a web server, and view the visualization directly from a Jupyter Notebook.

In analysing errors in POS tagging, we segment them into critical errors and non-critical errors. Critical errors are those that can change the semantic interpretation of an entire sentence, typically due to a assigning an entirely incorrect POS category to a word, for example a Plural Noun (NNS) incorrectly tagged as a Present Tense Verb (VBZ). This alteration in the semantics has a deleterious effect on all the subsequent steps in the NLP pipeline – e.g. Syntactic Parsing, Dependency Parsing, etc.

Non-critical errors, on the other hand, are those where the effect of the tagging error is local and have less significant consequences. In this study, we have not only seen the classes of POS tagging approaches but also developed and enhanced the module using algorithm. The developed taggers are enhancement of stochastic and rule based tagger to tackle online texts With this enhancement, both of the developed taggers are capable of handling unknown word and ambiguity which are two undeniable issues faced in any POS tagging application.

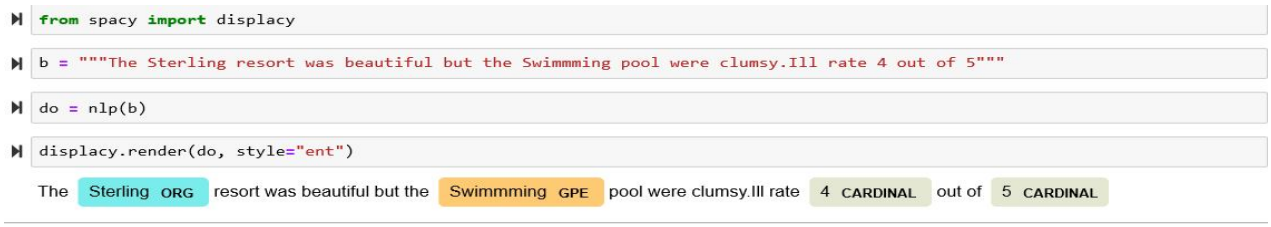

Fig. 6 Example review where important details are highlighted

VIII. CONCLUSIONS

The very thought of processing thousands of reviews makes us nauseous. Analysis of data is essential in this information world. People with the help of this data can make themselves equipped and also improve their performance from that of their previous experiences. In this project we have successfully established a review analysis system, which can be used to find the topic on which the review emphasis and make the admin gain useful insights from it. This recommendation system can also be used to other type of recommendation systems and also data that requires multi-label classifications.

ACKNOWLEDGMENT

I would like thank PSG College of Technology for supporting my work especially the faculty of the department Computer of Science and engineering and their staffs; students and my colleagues who help me in publishing my work.

REFERENCES

1. Rakesh Chandra Balabantaray. Sentiment Analysis of Movie Reviews using POS tags and Term Frequencies. IIIT Bhubaneswar, Volume 96– No.25, June 2014.
2. N. Cristianini and J. Shawe-Taylor. Introduction to Support Vector Machines. Cambridge University Press, 2000.
3. T. Joachims. Text categorization with support vector machines: learning with many relevant features. In Claire Ne´dellec and Ce´line Rouveirol, editors, Proceedings of ECML-98, 10th European Conference on Machine Learning, number 1398, pages 137–142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.
4. Sida Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In ACL.
5. Paul Rayson, Dawn Archer, Alistair Baron, Jonathan Culpeper and Nicholas Smith. Tagging the Bard: Evaluating the Accuracy of a Modern POS Tagger on Early Modern English Corpora. Researchgate.net publications, July 2007.
6. Christopher D. Manning. Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?. Departments of Linguistics and Computer Science Stanford University, LNCS, volume 6608, July 2010.
7. Eric Brill. A SIMPLE RULE-BASED PART OF SPEECH TAGGER. Department of Computer Science University of Pennsylvania Philadelphia May 2002.