Restaurant Review Dataset Based Sentimental Analysis with Machine Learning Approach

A Project Work Synopsis

Submitted in the partial fulfilment for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE WITH SPECIALIZATION IN

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Submitted by:

21BCS8803 M.L.K. SUBRAHMANYAM

21BCS8824 K. PRAKASH

20BCS6832 G.SAI PRASANTH

20BCS6380 T. BHAGEERATH

Under the Supervision of:

Prof. Siddharth Kumar

CHANDIGARH UNIVERSITY, GHARUAN, MOHALI - 140413,

PUNJAB

March, 2023

Abstract

Sentiment analysis of customer reviews has a crucial impact on a business's development strategy. Evolution of the internet in the past decade resulted in generation of voluminous data in all sectors. Due to these advents, the people have new ways of expressing their opinions about anything in the form of Google Reviews, Tweets, Blog Posts etc. Sentiment analysis deals with the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude toward a particular topic is positive, negative, or neutral. Knowing the opinion of customers is very important for any business. Hence, in this, we analyse the reviews given by the customers of the restaurant with help of machine learning classification algorithms. The Modern area of sentiment mining also called opinion mining. Researcher in the area of natural language processing (NLP), data mining, machine learning, Support Vector Machine (SVM) and test the method of sentiment analysis process. This problem can be addressed by an automated system called sentiment analysis and opinion mining that can analyse and extract the users view in the reviews.

Keywords— Sentiment Analysis, Category-Classification, Naïve Bayes Classifier, Logistic Regression, Support Vector Machine, Random Forest, Natural Language Processing (NLP), Restaurant Reviews Classification, and Machine Learning

Table of Contents

1. INTRODUCTION

Sentiment Analysis is the way toward deciding if a portion of writing is negative or positive. This piece of writing could be a tweet, review about a book, film, movie, restaurant and so on. The sentiment analysis is also known as opinion mining, in which the opinions, appraisals, emotions or attitude towards a topic, person or entity are analysed. The expressions can be classified as positive or negative. For example. "I really liked the garlic noodles of your restaurant "- this is a positive expression. The overall sentiment polarity shows a preference on service in the reviews, which might hint the customers to "self -select" the food they like. Natural language processing in artificial intelligence applications makes it easy to gather product reviews from a website and understand what consumers are actually saying as well as their sentiment in reference to a specific product. Companies with a large volume of reviews can actually understand them and use the data collected to recommend new products or services based on customer preferences.

It tends to be utilized to recognize the client or consumer's mentality towards a brand's crucial factors, for example, tone, context, emotion and so forth. These sorts of reviews are equally important for both the consumers and the betterment of the service. From consumers perspective having a view over that service from other consumers is useful for him to get an overall idea of the product. On the other hand, owners or service providers use sentiment analysis to have view about the acceptance of their products or to analyse customer satisfaction and suggestions. But as we can assume it is a very lengthy and time-consuming approach to go through that huge number of reviews and manually analyse the sentiment of those contents. Using sentiment analysis, the overall result about the

opinions and views can be obtained within seconds. It not only gives the owners an idea about the consumers, but it also gives them a better picture of how they stack up against their competitors' company. Restaurant reviews are still in the form of text, customer reviews are included in the text mining category, the results of these data will be classified into two values, positive or negative. for pre-processing review data such as remove stop word, remove punctuation done with the help of Python.

1.1 Problem Definition

To provide a Sentiment Analysis system for customers review classification, that may be helpful to analyse the information where opinions are highly unstructured and are either positive or negative.

1.2 Problem Overview

The major step involved in determining the sentiment of a text. In our approach, we have split the pre-processing part into three major steps.

The first step involves removing the punctuation in the sentences. All special characters like exclamatory mark and quotes are removed by designing appropriate regular expression. The resultant data would be containing only alphabetical characters. The second step involves removing the stop-words from the reviews. Stop-words are the words which are not used to express any emotion or sentiment but used as connectors or articles in the English language. This includes words like and, with, of, the. Natural language processing (NLP) techniques like Lexical analysis, syntactic analysis, semantic analysis, disclosure integration, and pragmatic analysis are applied on the dataset to identify and remove stop-words. The semantic analysis step generally removes the stop-

words like not as well. But, in opinion mining, the presence/absence of the word not plays an important role. For example, the review says the crust is not good. The removal of stop-words will result this sentence into crust good. Thus, a negative opinion is turned into positive. To avoid this problem, we have modified the semantic analysis step in NLP and made sure that such stop-words are not being removed in the process. The third step in pre-processing is to convert the original words to their root words. Root words are the words without prefix or suffix. For example, love is the root word for the words loving, loved, loves, etc. As we are interested only in actual opinion/sentiment rather than English grammar, such conversion eases the job. The Porter Stemmer algorithm is applied for converting all words in the dataset into root words.

1.3 Hardware Specification

Personal computer with keyboard and mouse maintained with uninterruptedpower supply.

- Processor: Intel® core™ i5
- Installed Memory (RAM): 8.00 GB

1.4 Software Specification

Operating System   : WINDOWS 7, 8.1,10,11
Coding language   : PYTHON
Web Browser   : GOOGLE CHROME

2. LITERATURE SURVEY

Sentiment analysis is used to learn about the customers' opinions about restaurant's services through their Twitter discussions. Naïve Bayes (NB) classification model is used for sentiment analysis to classify tweets into positive and negative. Even if NB is a simple, probability-based classification model, many researchers have used sentiment classification. In marketing and analysis, customer loyalty is a fundamental issue in terms of consumer success. As with hotel customer habits, they will pass on the mouth to mouth to other people when they are given excellent service. Text extraction or data retrieval is often done using analytical methods or manuals from the document collection store.

The research method will generate knowledge that can boost revenues and services from various text mining perspectives. Analyzes of sentiments are used to find views of a given object from the author . An opinion study on a commodity is a sentiment evaluation analysis. Sentiment analysis is based on the Natural Language Processing (NLP), the analysis of text and certain measured sections to delete or exclude unnecessary parts to interpret the pattern of the term negatively / positive. For sentiment analysis, the use of data mining algorithms has been extensive in the past. Now let's look at how powerful NB is and how widely it was employed as an essential classification data mining algorithm. NB classification is used for seismic and nuclear explosion detection.

2.1 Existing System

The content of user generated opinions in the social media such as face book, twitter, review sites, etc are growing in large volume. These opinions can be tapped and used as business intelligence for various uses such as marketing, prediction, etc. Generally, sentiment analysis is used for finding out the aptitude of the author

considering some topic. But in our social network sites not implemented Sentiment analysis. Some survey depends on the static sent word dataset to find the sentiment analysis. But we require finding a proper solution to find the polarity of the micro blogs.

2.2 Proposed System

We will collect the unstructured data through the text box. With that data covert the data to lower case and data is processed as follow. Pre-processing Before the feature extractor can use the reviews to build feature vector, the review text goes through pre-processing step where the following steps are taken. These steps convert plain text of the review into process able elements with more information added that can be utilized by feature extractor. For all these steps, third-party tools were used that were specialized to handle unique nature of review text.

Step 1: Tokenization

Tokenization is the process of converting text as a string into processable elements called tokens. In the context of a review, these elements can be words, emoticons, url links, hashtags or punctuations "an insanely awsum…." Text was broken into "an", "insanely", "awsum"…. These elements are often separated by separated by a space. On the other hand, hash tags with"#" preceding the tag needs to be retained since a word as a hash tag may have different sentiment value than a word used regularly in the text.

Step 2: Parts of Speech Tags

Parts of Speech (POS) tags are characteristics of a word in a sentence based on grammatical categories of words of language.

This information is essential for sentiment analysis as words may have different sentiment value depending on their POS tag. For example, word like "good" as a noun contains no sentiment whereas "good" as an adjective positive sentiment. each token extracted in the last step is assigned a POS

Step 3: Dependency Parsing

For our purposes, dependency parsing is extracting the relationship between words in a sentence. This can be useful in identifying relationship between "not" and "good" in phrases like "not really good" where the relationship is not always with the adjacent word.

CLASS	REVIEWS
positive	Had A Great Experience After so long. amazing Elite serving experience
negative	Service is very slow, Worst hospitality, Worst experience...
neutral	My first visit experience was great and over a period of time it changed to not bad. Reduction in quantity, timeliness also to an extent quality.
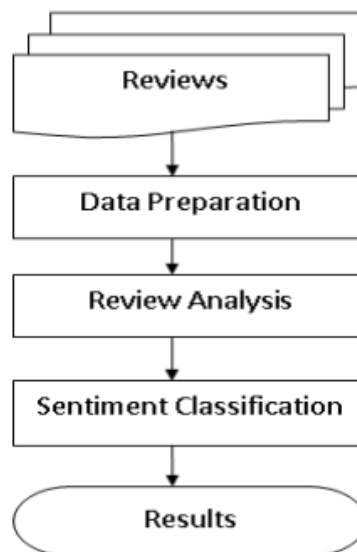
3. PROBLEM FORMULATION

Actually, most of businesses remain as failures due to lack of sufficient profits, lack of proper improvement measures. Specifically, the restaurant owners are facing a lot of problems and difficulties to improve their businesses due to lack of productivity in their business. Based on a report, the number of restaurants that have been shutting down in the first year are upto 60% and the restaurant businesses which are stopping their services within the first 5 years are about 80% in the world. So, this is becoming a universal problem now. There are various reasons involved in the failure of restaurant businesses, but serving low-quality food and lack of proper taste in the food items they are serving is a major one. To improve their business, they should know drawbacks of their food items and improve them by taking quality measures.

To resolve the problem of business loss which is due to various drawbacks of food items, a solution is to implement a "RESTAURANT REVIEW ANALYSIS SYSTEM". Each food item in the restaurant is assessed through the text review given by the customer where the text review is verified as a positive review or a negative review by the implementation of collection of classification models (Collective ML Model) in machine learning after handling textual data with natural language processing (to convert textual data to numerical data). And this data is stored into the database where each food item consists of number of customers, number of positive reviews, number of negative reviews, positive rate and negative rate. Now, the owner checks the least positively rated food item from the database, and takes necessary quality measurements which involve replacement of chefs or updating the ingredients used etc.

4. OBJECTIVES

The objective of this project is to conduct a comprehensive analysis of restaurant reviews using sentiment analysis techniques to identify the sentiment expressed by customers towards different aspects of the restaurant, such as food quality, service, ambiance, and value for money. The study aims to compare the performance of different sentiment analysis algorithms and evaluate their accuracy in detecting different levels of sentiment, such as positive and negative sentiments. Additionally, the research aims to investigate the impact of various factors, such as demographics, location, and restaurant type, on the sentiment expressed by customers in their reviews.

Furthermore, the report aims to provide actionable insights and recommendations to restaurant owners and managers based on the analysis of the sentiment expressed by customers. The report will outline strategies that can be implemented to improve customer satisfaction, enhance the overall customer experience, and boost business revenue. Overall, the objective of this project report is to demonstrate the value of sentiment analysis in understanding customer sentiments towards restaurants and provide a practical framework for restaurant owners and managers to leverage the insights gained from sentiment analysis to improve their business operations.

5. METHODOLOGY

This proposed work is to predict the text automatically based on the data set values stored by using the r tool. By using the training data set values, it is Possible to predict the text data using our classifier called naive bayes using algorithm.

The Fig I depicts the architecture of the proposed model used in the prediction of sentiment analysis. It consists of 3 steps

A. Data Collection

In this step data is taken out from Kaggle in a recognized format. Missing fields are evacuated in this process & thus the data is transformed. Sentiment Analysis can be considered a classification process. There are three main classification levels in sentiment analysis document-level, sentence-level, and aspect-level sentiment analysis. Level of document it aims to classify an opinion document which as a positive or negative opinion expression. It considers the full document as a basic information unit.

B. Data Pre-processing

The collected raw data of restaurant reviews consist of large number of attributes and there will be missing values. The reducing the attributes is required, extracting the attributes is also much essential. So, in order to meet importance of each variable or attributes "migrittr" algorithm is applied. Migrittr algorithm which selects the attributes based on predictor, here predictor considered restaurant review. Feature or Attribute extraction is done using migrittr algorithm. In detail steps working of migrittr algorithm. In Data cleaning once attributes are removed, filling the missing values, removing inconsistent data measuring the central tendency

for the attribute such as mean median, quartile is done. In data pre-process the data is cleaned and the extracted data before analysis. Non-textual contents and contents that are irrelevant for the analysis are identified and eliminated.

C. Sentiment Analysis

The reviews sources are mainly review sites. Sentiment analysis is not only applied on product reviews but can also applied on stock market, news articles, or political debates. In political debates for example, we could figure out people's opinions on a certain election candidates or political parties. The election results can also be from political posts. The sites like social media and micro

blogging sites are taken a very good source of information because many people share and discuss their opinions about positive and negative opinion freely.

D. Classification

The lexicon-based approach is to finding the opinion mining which is used to analyze or to predict the text. There are two methods in this approach. The dictionary- based approach which depends on finding opinion seed words, and then searches the dictionary of their synonyms and antonyms. The corpus-based approach begins with a seed list of opinion words, and then finds other opinion words in a large corpus to help in finding opinion words with context specific orientations. This could be done by using statistical or semantic methods.

Data mining has got two most frequent modelling goals — classification & prediction. Classification model classifies discrete, unordered values or data. In this prediction process, the classification techniques utilized are, naive bayes classifier.
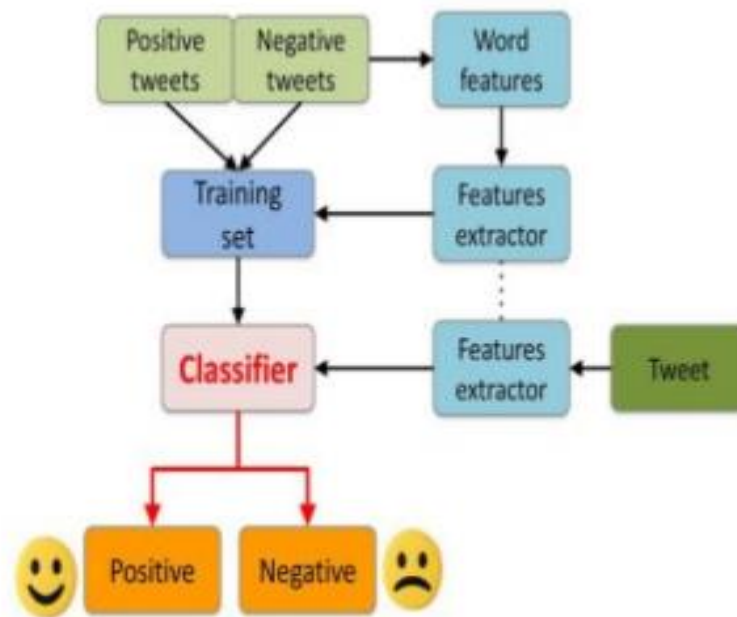
Fig1: Architecture of the proposed model

A. Naive Bayes

It is one of the popular classification techniques of algorithms used in data mining. It is a probability classifier. It links the attributes mutually & is dependent on the number of parameters, The principle here is that the variables provided are independent. It generates accurate results with appropriate calculation & provides fast results. It is on Bayes theorem & the formula is,

$$P \text{ (label | features)} = \frac{P(\text{label}) * P \text{ (features | label)}}{P(\text{features})}$$

7.CONCLUSION

The proposed work starting from the analysis of different studies provided in the literature, provides a classification of sentiment classification approaches with to features/techniques and advantages limitations, tools for sentiment analysis with respect to the different techniques used for sentiment analysis.

The sentiment classification approaches can be classified in machine learning, lexicon based and hybrid approach. The machine learning approach is used for predicting the sentiments based on trained and test data. In our lexicon-based approach does not need any prior training in order to mine the data.

8. TENTATIVE CHAPTER PLAN FOR THE PROPOSED WORK

CHAPTER 1: INTRODUCTION

This chapter will cover the overview of Sentiment Analysis.

CHAPTER 2: LITERATURE REVIEW

This chapter include the literature available for Sentiment Analysis. The findings of the researchers will be highlighted which will become basis of current implementation.

CHAPTER 3: OBJECTIVE

This chapter will provide introduction to the concepts which are necessary to understand the proposed system

CHAPTER 4: METHODOLOGIES

This chapter will cover the technical details of the proposed approach.

CHAPTER 5: EXPERIMENTAL SETUP

This chapter will provide information about the subject system and tools used for evaluation of proposed method.

CHAPTER 6: CONCLUSION AND FUTURE SCOPE

The major finding of the work will be presented in this chapter. Also, directions for extending the current study will be discussed.

REFERENCES

[1] M. Govindarajan, "Sentiment Analysis of Restaurant Reviews Using Hybrid Classification Model," IRF International Conference, 2014.

[2] D. V. N. Devi, C.K. Kumar and S. Prasasd "A feature Based Approach for Sentiment Analysis by Using Support Vector Machine".

[3] Y. Woldemariam, "Sentiment analysis in a cross-media analysis framework," Hangzhou, 2016.

[4] E. Boiv and M. F. Moens, "A machine learning approach to sentiment analysis in multilingual Web texts," Belgium: Springer- Information Retrieval Journals, 2008.

[5] A.B. Pawar, M.A. Jawale and D.N. Kyatanavar, "Fundamentals of Sentiment Analysis: Concepts and Methodology", An Environment of Computation Intelligence, Springer International Publishing Switzerland 2016, pp. 25-35.

[6] Peter Koncz and Jan Paralic, "An approach to feature selection for sentiment analysis", 15th International Conference on Intelligent Engineering Systems, June 23 –25, 2011, Poprad, Slovakia.

[7] S.Vidhya, D.Asir Antony Gnana Singh and E.Jebamalar Leavline, "Feature Extraction for Document Classification", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 4, Special Issue 6

[8] B.Pang and L.Lee, "Opinion mining and sentiment analysis," vol. 3, no. 1-2, 2008.

[9] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up sentiment classification using machine learning techniquest," in In ACL Conference on Empirical Methods in Natural Language Processing, 2010, pp. 354-368.

[10] C. Musto, G. Semeraro, and M. Polignano, "A comparison of lexicon-based approaches for sentiment analysis of microblog posts," on 8th International Workshop on Information Filtering and Retrieval, 2014.