ASPECT BASED RESTAURANT REVIEW SUMMARIZATION SYSTEM USING MACHINE LEARNING

Ms .S.Sangeetha mariammal, M.E, Associate professor,CIT

G. Abirami, S .Roshan Prakya, S.S. Sneha Priyadarshini, A.Surekha

B.E, Dept. of Computer Science Engineering, CIT, Coimbatore

ABSTRACT- The most convenient way of knowing about restaurant before ordering food is through user reviews. Mostly reviews are too long to read and time consuming. Representing the reviews to the user in a summarized way saves time and provides clarity. In this proposed system the summarization of reviews based on the common aspects are implemented using aspect based sentimental analysis, opinion mining and machine learning. Sentiment analysis refers to the use of natural language processing and text analysis to systematically identify, extract, and quantify the subjective information. Support Vector Machines (SVM) machine learning classification technique is used to categorize sentences based on aspects and it increases the accuracy. This proposed system aims to produce the best review summarization model that facilitates the user for deciding the restaurant. This model can be used for any product review summarization.

KEYWORDS: Aspect based sentimental analysis, polarity, Support Vector Machines.

1. INTRODUCTION

Sentiment analysis (also known as opinion mining or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information.. The existing system shortens the reviews provided by the users because it takes a lot of time to read all the reviews. The existing system provides suggestion by using the description given by the management which may or may not be true. The proposed system also shortens the reviews with increased speed along with clustering techniques. The proposed system is to build a system that provides gist of reviews which is more reliable than ratings which are less realistic and not accurate enough.

MACHINE LEARNING:

Machine learning is a subfield of artificial intelligence that allows machines to access data themselves, learn from this data, and perform tasks. This is done through learning algorithms and statistical models.

Unsupervised learning is where you only have input data (X) and no corresponding output variables. The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

Unsupervised learning problems can be further grouped into clustering and association problems.

- Clustering: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
- Association: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

$$Y = f(X)$$

The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

Supervised learning problems can be further grouped into regression and classification problems.

- Regression: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.
- Classification: A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”. Types of some classification algorithms in machine learning are Logistic Regression Nearest Neighbor, Support Vector Machine and Random Forest Classification

OPINION MINING:

Opinion mining also known as Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials.

NLP:

Natural language processing (NLP) is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.

NLTK:

Natural Language Tool Kit(NLTK) is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as Word Net, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

TEXT PREPROCESSING:

To pre-process your text simply means to bring your text into a form that is ***predictable*** and ***analyzable*** for your task.

Steps in pre-processing:

- Lower case to upper case.
- Remove stop words
- Lemmatization
- Convert numbers to words
- Remove punctuation and blank spaces

ASPECTS BASED SENTIMENT ANALYSIS:

The next step is classifying reviews into some of the predefined categories. These categories are nothing but the aspects that are frequently recurring in the review data set. The big difference between sentiment analysis and aspect-based sentiment analysis is that the former only detects the sentiment of an overall text, while the latter analyzes each text to identify various aspects and determine the corresponding sentiment for each one. In other words, instead of classifying the overall sentiment of a text into positive or negative, aspect-based analysis allows us to associate specific sentiments with different aspects of a product or service. The results are more detailed, interesting and accurate because aspect-based analysis looks more closely at the information behind a text.

2. LITERATURE SURVEY

A. A Summarisation Tool for Hotel Reviews:

Summarization is a way to shorten all the reviews without change the actual meaning of the sentence. In order to summarise the hotel reviews, a method named Featured Noun Pairing has been chosen. This method associates features (nouns) and adjectives to represent a whole review sentence. Besides that, this system will also provide a function to analyze textual data where involving natural language processing tasks. The tasks are tokenization, stop words removal, and others

B. Aspect based Sentiment Oriented Summarization of Hotel Reviews

This study analyzes the hotel reviews and gives information that ratings might overlook. The reviews and metadata are crawled from website and classified into predefined classes as per some of the common aspects.
Then Topic modeling technique (LDA) is applied to identify hidden information and aspects, followed by sentiment analysis on classified sentences and summarization.

C. A Novel Automatic Sentiment Summarization from Aspect-based Customer Reviews :

In this system aspect-based representation used to represent ranked knowledge on aspect opinion calculated by using frequencies, polarity, and opinion strength. The second phase is the review summary generation used to automatically produce review summary by ranking aspect based on information of the aspect.

D. A New Profile Learning Model for Recommendation System based on Machine Learning Technique:

The main contribution of this paper is to introduce a new user profile learning model to promote the recommendation accuracy of vertical recommendation systems. The proposed profile learning model employs the vertical classifier that has been used in multi classification module of the Intelligent.

3. TECHNIQUES AND TOOLS

POLARITY:

Polarity in sentiment analysis refers to identifying *sentiment orientation* (positive, neutral, and negative) in written or spoken language. Other types of sentiment analysis include fine-grained sentiment analysis which provides more precision in the level of polarity (e.g. very positive, positive, neutral, negative, and very negative) and emotion analysis which aims to surprise identify emotions in expressions (e.g. happiness, sadness, frustration**,,** etc).Language can contain expressions that are *objective* or *subjective*. Objective expressions are facts. Subjective expressions are opinions that describe people's feelings towards a specific subject or topic.

WEB SCRAPING:

Web scraping is a term used to describe the use of a program or algorithm to extract and process large amounts of data from the web. Whether you are a data scientist, engineer, or anybody who analyzes large amounts of datasets, the ability to scrape data from the web is a useful skill to have. Let's say you find data from the web, and there is no direct way to download it, web scraping using Python is a skill you can use to extract the data into a useful form that can be imported

- Data extraction from the web using Python's Beautiful Soup module
- Data manipulation and cleaning using Python's Pandas library

BEAUTIFULSOUP:

Beautiful Soup is another beautiful python module which aids scraping the data required from html/xmls via tags. With beautiful you can scrape almost everything because it aids different methods like searching via tags, finding all links, etc.

PANDAS:

Pandas is the most popular python library that is used for data analysis. It provides highly optimized performance with back-end source code is purely written in *C* or *Python*

SVM:

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.SVM chooses the extreme points/vectors that help in creating the hyper plane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyper plane.

TYPES OF SVM:

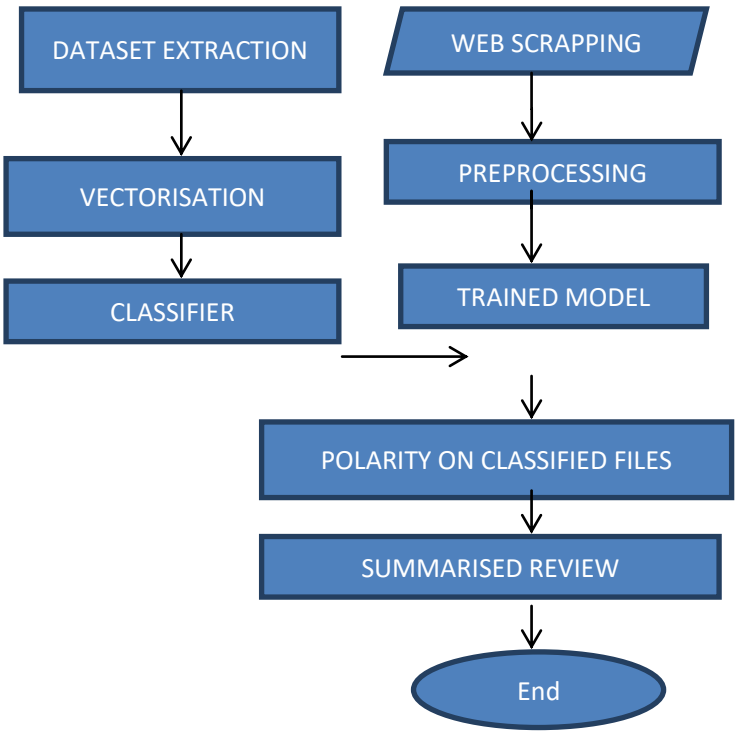
SVM can be of two types:

- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

SKLEARN:

Scikit-learn (also known as **sklearn**) is a free software machine learning library for the Python programming language.It features various regression and clustering algorithms including support vector machines, random forests, gradient boosting, *k*-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy

4. PROPOSED SYSTEM



- The data is extracted from review website using Web scrapping and it is preprocessed.
- Then the necessary data for training the classifier model is vectorized and classified using SVM classifier.
- After classifying the data polarity of the aspects are identified and hence the summarized review is obtained.

5. RESULT

MODULE 1 : Web Scrapping and Pre-processing

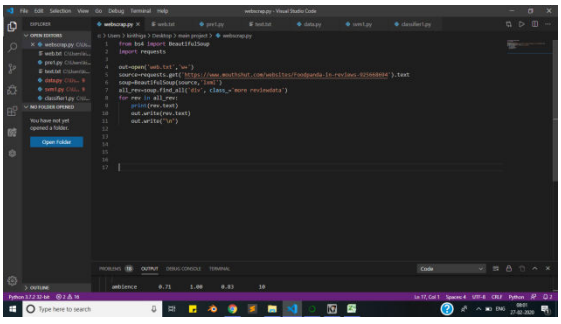


Fig: 5.1

The necessary review dataset is extracted from the html page source code using Web scrapping in Fig 5.1

Web scrapped text file:

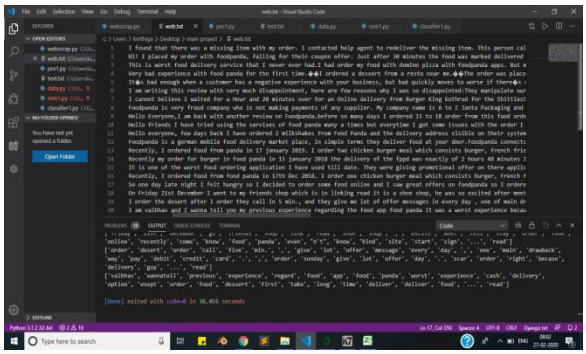


Fig: 5.2

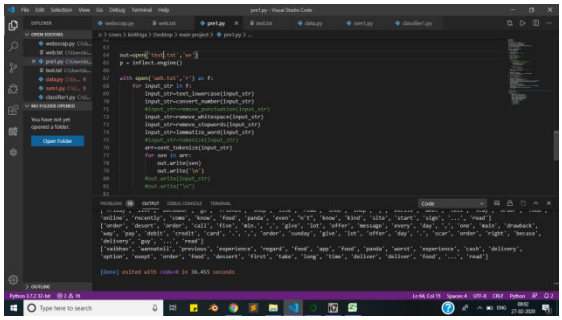


Fig: 5.3

The extracted dataset is pre-processed by converting into lowercase, removing whitespace, removing stopwords, lemmatizing and tokenizing the words.

MODULE 2: Preprocessed Text File

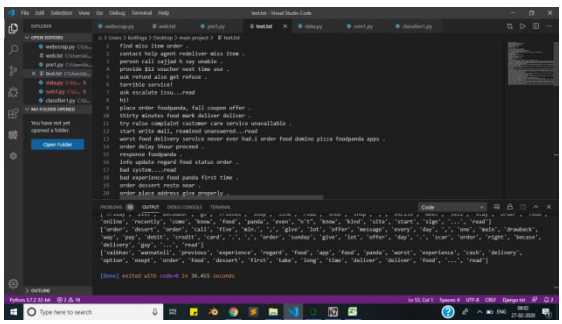


Fig: 5.4

MODULE 3: SVM classifier

Training and Testing Dataset

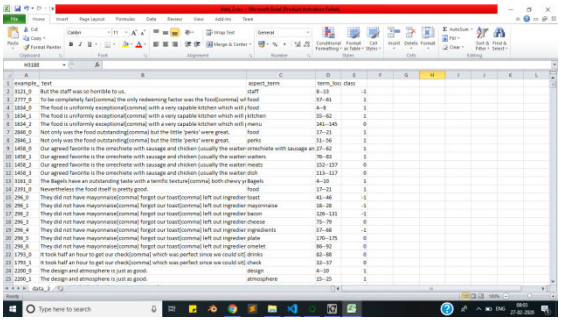


Fig: 5.5

Accuracy of SVM classifier for the given review dataset in fig 5.6

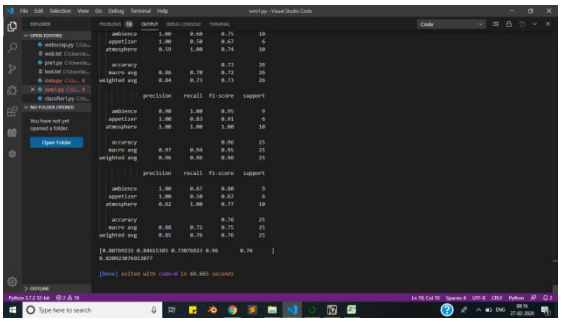


Fig: 5.6

CONCLUSION

Thus this system provides suggestion to the users using the reviews provided by the customers which are shortened using the mining techniques hence time saving.

REFERENCES:

i. A Summarisation Tool for Hotel Reviews 2018 N.H.Abd Rahim M.A.F Hasnan

ii. Aspect based Sentiment Oriented Summarization of Hotel Reviews.
2017 Nadeem Akhtara, Nashez Zubaira, Abhishek Kumara, Tameem Ahmada *aD*

iii. A Novel Automatic Sentiment Summarization from Aspect-based Customer Reviews
2018 Tuan Anh Tran JaruneeDuangsuwan Wiphada Wettayaprasit

iv. A New Profile Learning Model for Recommendation System based on Machine Learning Technique
2016 Shereen H Ali, Ali I El Desouky and Ahmed I Saleh