



**Project Report**

**On**

Restaurant Review Sentiment Analysis



Submitted in partial fulfillment for the award of

**Post Graduate Diploma in Big Data Analytics-(DBDA)**

**From**

**Know-IT (Pune)**

Guided by:

**Mr. Amey Manjrekar**

Presented by:

Akriti Gupta 220343025004

Gayatri Pandit 220343025014

Prathamesh Pachpute 220343025030

Prayag Kesarwani 220343025020



**Centre of Development of Advanced Computing (C-DAC), Pune**

**CERTIFICATE**

**TO WHOMSOEVER IT MAY CONCERN**

**This is to certify that**

**Ms. Akriti Gupta PRN: 220343025004**

**Ms. Gayatri Pandit PRN: 220343025014**

**Mr. Prathamesh Pachpute PRN: 220343025030**

**Mr. Prayag Kesarwani PRN: 220343025020**

**Have successfully completed their project on**

Restaurant Review Sentiment Analysis

**Under the guidance of**

**Mr. Amey Manjrekar**

**Project Guide Project Supervisor**



**ACKNOWLEDGEMENT**

This project **“Restaurant Review Sentiment Analysis”** was a great learning experience for us and we are submitting this work to CDAC Know-IT. We all are very glad to mention the name of **Mr. Amey Manjrekar** for his valuable guidance to work on this project. His guidance and support helped us to overcome various obstacles and intricacies during the course of project work.

We are highly grateful to **Mr. Vaibhav Inamdar** (Manager (Know-IT), C- DAC, for his guidance and support whenever necessary while doing this course Post Graduate Diploma in **Big Data Analytics** (PG-DBDA) through C-DAC ACTS, Pune.

Our most heartfelt thanks goes to **Mrs. Bakul Joshi** (Course Co-Ordinator, **PG-DBDA**) who gave all the required support and kind co-ordination to provide all the necessities like required hardware, internet facility and extra Lab hours to complete the project and throughout the course up to the last day here in C-DAC Know-IT, Pune.

**From:**

**Akriti Gupta PRN: 220343025004**

**Gayatri Pandit PRN: 220343025014**

**Prathamesh Pachpute PRN: 220343025030**

**Prayag Kesarwani PRN: 220343025020**

**CONTENTS**

**1.INTRODUCTION**

1.1 Document Purpose

1.2 Project Background

* 1. Problem Statement

1.4 Objective

**2. FUNCTIONAL REQUIREMENTS OVERVIEW**

* 1. Technologies Used

2.2 Apache Spark

2.3 MongoDB

2.4 Python

* 1. Power BI

**3. SYSTEM REQUIREMENTS**

* 1. Hardware Requirements
  2. Software Requirements

**4. PROJECT FLOW**

**5. METHODOLOGY**

5.1 Data Used

5.2 Data Dumping

**6. RESULTS AND ANALYSIS**

**7. FUTURE SCOPE**

**8. CONCLUSION**

**9. LINKS AND REFERENCES** 



**ABSTRACT**

After COVID, there was a sudden surge in demand for online food ordering which eventually has increased the value of reviews and ratings given by the customers for a particular restaurant. The ability to analyze and extract information from customers has great value to a business that can leverage the data to continuously improve.

This project comprises of the analysis to analyze sentiments based on customer reviews for restaurant business. The project focuses on using open source restaurant data, processed by Spark and stored in Mongo DB on which we are implementing Deep Learning for sentiment analysis using NLP (NLTK, Text Blob), Machine learning algorithms (Random Forest, Logistic Regression, etc. ), Data Visualization tools and techniques(Power BI).

**INTRODUCTION**

**Document Purpose**

The purpose of this document is to build a system to analyze Restaurant Reviews Sentiment Analysis. The scope of this document is to define the functional and non-functional requirements, business constraints requirements.

**Project Background**

The project converges on using open source data to analyze the sentiments of the customers for the restaurants. On the basis of customer reviews the project indulges model building for improvement of the restaurant business. The project uses text data of different reviews and applies various suitable models which help in the analysis of sentiments on the basis of customer reviews. The project also comprise of recommending the restaurants to new customers. The project consists of in-depth analysis and visualization on the data for achieving improvement in business.

**Problem Statement**

To perform sentiment analysis on customer reviews and to build a recommendation system using customer reviews to recommend restaurants.

**Objective**

The objective of sentiment analysis is to accurately extract people's opinions from a large number of unstructured review texts and classifying them into sentiment classes,

i.e., positive, negative, or neutral.



**FUNCTIONAL REQUIREMENTS OVERVIEW**

**Technologies used:**

1. **Python:**

Python is open source, interpreted, high level language and provides great approach for object-oriented programming. It is one of the best language used by data scientist for various data science projects/application. Python provide great functionality to deal with mathematics, statistics and scientific function.

Python is commonly used for developing websites and software, task automation, data analysis, and data visualization.

**2)Apache Spark:**

Apache Spark is a multi-language engine for executing data engineering, data science and machine learning on single-node machine or clusters. We performed ETL on raw data or use in data processing pipeline. Replacing incorrect and inappropriate text in dataset. Removing duplicated or missing data.

1. **MongoDB:**

MongoDB provides a mechanism to store and retrieve data in relaxed consistency model with advantages like horizontal scaling, higher availability and faster access. MongoDB (from humongous) is reinventing data management and powering Big Data as the world's fastest-growing database.

**4)Machine Learning:**

Machine learning is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. Machine learning is important because it gives enterprises a view of trends in customer behavior and business operational patterns, as well as supports the development of new products. Text analysis (TA) is a machine learning technique used to automatically extract valuable insights from unstructured text data.

**5)Power BI:**

Power BI Desktop is the crucial component of Microsoft Business Intelligence. It enables a user to import, analyze, and create reports out of complex information, streaming in from multiple data sources. The free version of Power BI allows the user to analyze data, create live reports, and publish them on the Internet.



**SYSTEM REQUIREMENTS**

**Hardware Requirements:**

* + Platform – Windows 64 bit
  + RAM –At least 8 GB of RAM
  + Peripheral Devices – Mouse, Keyboard
  + Hard Disk- 500 GB or above
  + A Stable Internet Connection

**Software Requirements with Versions:**

* Apache Spark- Version 3.3.0
* Java- Version java 11.0.15 / java 11.0.16
* Mongo DB- Version Mongo DB shell version v5.0.5 or above
* Python IDE- Version Python 3.10.7 or above
* Microsoft Power BI Desktop- Version: 2.106.883.0 64-bit



**PROJECT FLOWCHART**



**METHODOLOGY**

**Data Used**

The data used in this project was Open-source Restaurant review data of 3.35MB Size.

The data was retrieved from GitHub, in CSV/text format. Total count of record in original data is 10001.

**Data Cleaning**

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from dataset by removing null values, symbols, numbers from reviews, rating and restaurant column. In dataset and convert into correct data type using apache Spark.

**Data Dumping**

Data was dumped into MongoDB using Spark . MongoDB is a non-relational document database that provides support for JSON-like storage. Total three datasets are loaded in

MongoDB .

Raw\_dataframe:[‘Restaurant’,’Reviewer’,’Rating’,’Review’,’Metadata’,’Time’,’Pictures’]

Dataframe\_1: ['Restaurant', 'Review']

Dataframe\_2: ['Restaurant', 'Reviewer', 'Rating']



**RESULTS AND ANALYSIS**

**1.** The results show the calculation of polarity of reviews for restaurant business from which it is noticed that 70% reviews are positive and 30% reviews are negative.

**2.** The use of VADER (Valence Aware Dictionary for Sentiment Reasoning) results into classification of sentiments into these categories as positive, negative and neutral.

**3.** Word Cloud is depicted with the help of reviews which contain most frequently used words.

**4**.The cleaned reviews are used to perform recommendation of restaurant names on the basis of ratings.

**Machine Learning Models and Algorithm**

Based on data, various Machine learning and Deep learning Algorithms are applied

**Important Algorithms/ Libraries applied on Dataframe\_1 :**

TextBlob, NLTK(Natural Language Toolkit), WordCloud, Spacy, Sentiment Intensity Analyser, VADERS(Valance Aware Dictionary for Sentiment Reasoning),TF-IDF(Term Frequency Inverse Document Frequency)Vectorizer, lightgbm, scikit-learn, Collaborative filtering.

**Important Algorithms/ Libraries applied on Dataframe\_2 :**

Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest, Logistic Regression.

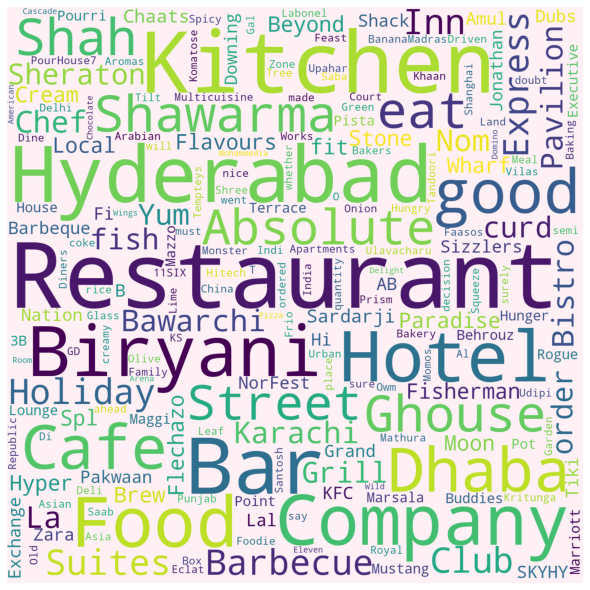
**Algorithms Accuracy**

1. Support Vector Machine (SVM) 0.754432
2. K-Nearest Neighbor (KNN) 0.699899

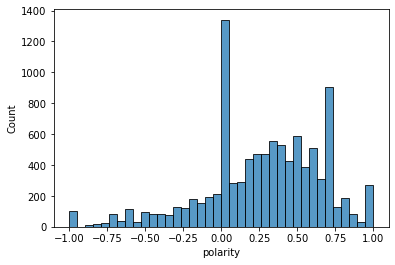
1. Random Forest 0.698895

1. Logistic Regression 0.596930

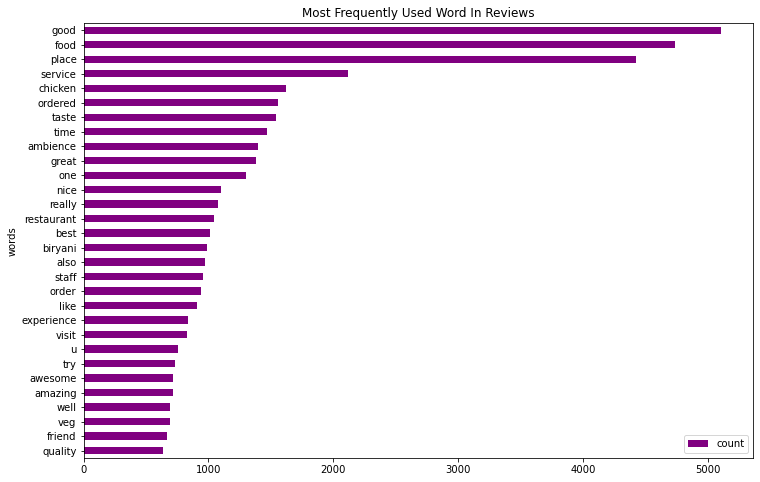




Word cloud with restaurant names







Power BI Charts Link: <https://drive.google.com/file/d/1GBJn7ZBU8JcrrRJ2-rjGm840aLI99mSj/view?usp=sharing>

**FUTURE SCOPE**

This project has a lot of strength, the analysis carried out is from business perspective and can extend the functionality and overview of the project by creating the application from user perspective.

In order to increase the performance and speed of analysis, the data can be dumped on cloud so that there is no requirement to load data every time the analysis is carried out.

The analysis performed considers only restaurant domain of the businesses and user review dataset. The future scope can consider different domain and can consider different sub datasets like tips and check-ins.

Future work includes building a module by which all these trends and insights are communicated to business in real time so that businesses can get faster feedback.

Sentiment analysis methods proved to perform well for classifying sentiments of restaurant business reviews by taking into account the rating given by the users. Further, the systems accuracy could still be improved by exploring the usage of bigrams or trigrams, word chunks, or even part-of speech (POS) as features in order to distinguish between the same word features that are used as different POS.



**CONCLUSION**

The designed system fulfills the objective of doing sentiment analysis of reviews of restaurant business and provide important insights and information to the business owner. The system also recommends the users to choose restaurant based on reviews and ratings. The algorithms such as TextBlob and NLTK are used to process the text data that will provide great strength to the sentiment analysis of restaurant dataset. We believe that our system will definitely add more value to the business that will help to grow their business.

**LINKS AND REFERENCES**

* <https://raw.githubusercontent.com/gayatripandit006/CDAC--DBDA-Project/main/Restaurant_Reviews_Raw.txt>
* <https://towardsdatascience.com/data-visualization-using-matplotlib-16f1aae5ce70>
* <https://textblob.readthedocs.io/en/dev/quickstart.html>
* <https://docs.python.org/3/>