Analysis of Reviews using Big Data Technologies

The Team

Team Members	Student Id	Email Addresses	Expected Roles
Bhavana Kurra	110294451	bhavana.kurra@ucdenver.edu	Building a Machine Learning Model
Vijayasimha Bheemireddy	109965858	vijayasimha.bheemireddy@ucdenver.edu	Analyzing and Creating Visualizations
Praveen Vatambeti	110312985	praveen.vatambeti@ucdenver.edu	Building a Machine Learning Model
Navya Mogili	110315073	navya.mogili@ucdenver.edu	Extraction and Processing of Data

Problem Statement and Background

As the number of restaurants are steadily increasing over the years, the competition in gourmet sector has been increasing. In order to keep up with the competition and maintain the latest market trends one has to analyze and make key decisions. Analyzing reviews and deriving meaningful insights helps businesses to maintain their existing standards and gain more traction from the customers. These reviews can be further classified either into positive or negative statements or can be used for understanding the emotions of the reviewed customer.

Description regarding the Data Source

The dataset we intend to use is from kaggle (https://www.kaggle.com/yelp-dataset/yelp-dataset?select=yelp_academic_dataset_review.json) and Yelp official website (https://www.yelp.com/dataset) which consists of 6 different json datasets that includes the data of businesses across 8 metropolitan areas in the USA and Canada. Also it has the data of the customers and their reviews. The datasets are well structured json documents with proper entity relationships/mappings. The statistics of the dataset is represented in the below table:

Parameter	Count
Reviews	8,635,403
Businesses	160,585
Pictures	200,000
Metropolitan areas	8

The dataset should be pre-processed before usage that includes various steps such as Cleaning, Integration of Data, Removal of Redundant and sparse values.

Goals of the project

The primary goal of this project is to extract, transform, store and analyze the customer review data such that we can derive insights from them while classifying the reviews. The following are the success metrics of this project –

1. Extracting and storing the data efficiently on the Hadoop cluster
2. Accurately classifying the reviews

Description of Big Data Systems and Tools that will be used

In this project we intend to use Hadoop, Map Reduce, HBase, Pig, MongoDB technologies for storing and processing the data. Also we plan to use Mahout, Tensor flow or Keras libraries to build Machine Learning models over the data stored and Tableau for visualizations.

System Architecture and Conclusion

The key components of our system architecture are :

- Data sources
- Data Storage Systems (Big Data Systems such as Hadoop Clusters)
- Exploratory Data Analysis and Visualization Systems
- Model Building Systems (Classification models)
- Analytics and Reporting Systems

Using the above mentioned components, deriving the insights and being successfully able to classify the reviews is the purpose of our project.