**Analysis of Customer Reviews using Big Data**

**A Project Report**

**Submitted by**

**Pavansai Pottimuthi (811235129,** [**ppotimu@kent.edu**](mailto:ppotimu@kent.edu)**)**

**Shirish Kumar Tallapelly (811229216,** [**stallape@kent.edu**](mailto:stallape@kent.edu)**)**

**Shiva Kumar Peddapuram (811235874,** [**speddapu@kent.edu**](mailto:speddapu@kent.edu)**)**

**1. Introduction**

**Motivation examples of this project:**

* Nowadays, we usually buy products online. To buy an item we take a lot of parameters into consideration like shirt color, ratings, reviews for that specific product, comments, price, and lot more. In this growing era, all brands are trying to get the digital presence in different forms like maintaining their own social media pages (selling their products), e-commerce sites, and official websites. All the consumers or the buyer of that specific product gives their feedback in different firms like comments, ratings, and reviews.
* So, what we think of it is the most importance for all kinds of companies to understand their customer impression in digital platforms as early as possible. If they learn positive, it may help them to publish their products in that model. If they get any negative response, then they can work upon those traits and then improve their products.
* Building automated bots or software can undergo all the reviews and then analyze them for that companies, this process is known as sentimental analysis. For this type of software, we use different machine learning algorithms to develop the model so that it can work on its own.

**2. Project Description**

* The first challenge is integrating all the customer’s data from various sources, thus collecting datasets from online sources helps us to analyze data.
* To get a quick overview of the data set we use the dataframe.info () function. Python is an excellent language for data analysis, due to the solid ecosystem of data-centric Python tools.
* The second issue is cleaning all the null values and converting the data into upper or lower case, removing emojis and all unwanted data.
* Visualization of data by using sentimental intensity analyzer thus converting data from text to vectors into polarity scores and plotting the data word cloud.
* The main issue is regarding the privacy and protection of the user data.

**Sentimental Analysis:**

* Sentimental Analysis is a technique which helps us to identify the emotions from the raw texts.
* Sentimental Analysis plays crucial role in the domain of NLP (natural language processing).
* Sentimental analysis plays key role in most of the E-commerce like: Amazon, Flipkart, Shopify.
* In general, Sentimental Analysis is also referred to as opinion mining, that uses techniques like machine Learning Algorithms, different types of Statistical Methods and NLP.
* NLP is used in recognising and extraction of useful and subjective information from the raw/cleaned textual data. Like Comments, emotions hidden in th text.
* For sentiment categorization, we have five different machine learning classifiers: Nave Bayes, K-Nearest Neighbour, Support Vector Machine, Logistic Regression, and Random Forest

**Sentimental Score:**

* We use sentimental score to define the emotions of the texts.
* Sentiment score usually lies in between -1, 0, 1
* -1 defines as Negative.
* Here Negative in terms of sentimental analysis means the text on which we are working has negative impact.
* 0 represents as neutral.
* Here Neutral Sentimental Score Means it has no impact on the research we are working on it.
* 1 represents as positive.
* Positive Sentimental Score means it shows positive text is found out in our text data.

**The workload distribution for each member in our team:**

* Initially we are planning to collect data from various sources.
* After gathering the data, we are going to merge all the data into a single file.
* Then we are going to clean the data on the file which has been merged in the earlier step so that we can reduce all the unnecessary values.
* Here cleaning the data means removing the null values unnecessary data from the dataset we have finalized.
* Some sort of examples related to unnecessary data means (in the timestamp we only consider only time in HH:MM)
* Then in further steps we are going to work on the sentimental analysis.
* Data cleaning and Data preparation in this step removes unwanted data.
* Plotting the word cloud by using polarity scores, thus analyzing different attributes in the dataset.
* Performing different types of visualisations on the sentimental analysis on the textual dataset that has been gathered.
* Visualization of our data with different kinds of charts, graphs, and checking sentimental score.

**3. Background**

**Related papers:**

Nandal in this paper classified amazon product reviews for sentiment analysis using SVM (Support Vector Machine) Tool. The study examined how words can shift in meaning depending on the context in which they are used, and how this impacts the overall evaluation of a product and its specific features.

Humera Shaziya classified movie reviews for sentiment analysis using WEKA Tool. They enhanced the earlier work done in sentiment categorization which analyzes opinions which express either positive or negative sentiment.

Ahmad Kamal designed an opinion mining framework that facilitates objectivity or subjectivity analysis, feature extraction and review summarizing. He used a supervised machine learning approach for subjectivity and objectivity classification of reviews. The various techniques used by him were Naive Bayes, Decision Tree, Multi-layer Perception and Bagging. He also improved mining performance by preventing irrelevant extraction and noise.

To estimate the semantic orientation polarity and its intensity for phrases, which serves as a foundation for sentiment-based computing, Orestes Appel used natural language processing (NLP) fundamental techniques, a sentiment lexicon improved with help from SentiWordNet, and fuzzy sets. Three different datasets are subjected to the suggested hybrid method, and the outcomes are contrasted with those attained by applying the Maximum Entropy and Naive Bayes techniques.

**Required hardware:**

* **Required Technical Knowledge: Python** Programming Language (Object-Oriented programming), Sentiment analysis, Natural Language Processing (for training the model).
* Integrated Development Environment PyCharm and Google Colob.
* Different Python libraries are used to analyze the data (clean unnecessary data).
* **Dataset**: Here dataset used in this project is Realtime customer data collected from sources.
* **Required hardware:** System with 64-bit ROM and 8GB RAM.
* **Input:** customer data with reviews and ratings on product.
* **Output:** Use qualitative and quantitative prediction models to forecast the Customer Review Analysis.

**4. Problem Definition**

The challenge is to create a platform that compiles all of the important indicators for a company, such as the most current reviews, overall rating, sentimental distribution, trending keywords, etc. Many evaluations, ideas, and complaints are left on a company's website by customers. Reading and understanding all of this takes a lot of physical labor, money, and time.

There are problems while taking the dataset, all the reviews will not be in the same language, so we must change the data into machine readable. There are some companies that ask media managers to give false reviews for their benefits thus we cannot analyze data properly. Thus, in cleaning and preparation step we change the data. By analyzing the data, companies can know where they are lagging and how to improve their products and how they can change their products to analyze whom they must target.