**BUSINESS ANALYTICS AND BUSINESS INTELLIGENCE**

**CAPSTONE PROJECT | GROUP 1**

**PROJECT SYNOPSIS ON SENTIMENT ANALYSIS OF DIFFERENT PRODUCTS**

**Under Guidance of Ms Anjana Agrawal**

**Under Taken By**

Abhilash R

Angela Susan Mathews

Ashwin. S

Deepthi H

Sajin Madhavan S

Sneha devadass

**INTRODUCTION:**

Advancement in web technology and their growth have created path for generation of huge size of data in the web for internet users. Social networking sites like Twitter, Facebook, Google+ have quickly attained popularity as they permit public to share and express their views about topics, make discussion with diverse communities, or post messages across the globe.

This project focuses mainly on sentiment analysis performed on combinations of data scraped from various sources like E-commerce websites and other social media websites which helps us to analyze the review comments of the products where opinions are highly unstructured, heterogeneous and are either positive, negative or neutral.

**OBJECTIVE & SCOPE :**

Generally, customers prefer to check for the reviews of a product before buying it either online or in stores. Rather than checking it in multiple sites, why not to check in a single instance / app which rates the products using sentiment analysis, based on the review comments which has been gathered from multiple social medias( twitter,youtube, quora,etc..) as well as from multiple E-commerce websites.

Our aim is to provide people a platform to take a decision on purchase of products based on reliable reviews and review ratings given by a larger audience.

**ANALYTIC APPROACH:**

Sentiment analysis and opinion Mining are techniques which helps in finding opinions in text, detecting emotions, which has become a hotspot in Natural Language Processing and Information Retrieval.

Considering that Sentiment words and phrases are the main indicator of sentiment classification, Classifying the sentiment of texts based on dictionary defines the polarity

confidence of the words. It is commonly known as classifying subjective opinion into

positive and negative. Existing Polarity detection(PD) methods may be

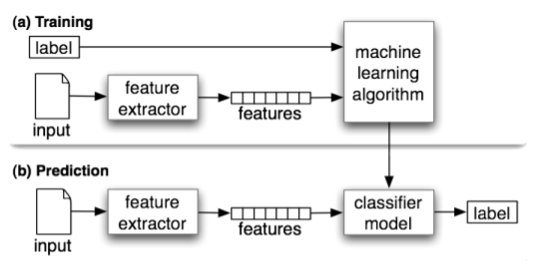
categorized into three main categories: Supervised, Unsupervised and Hybrid.

There are also other methods that takes different machine learning approaches and

build text classifies such as decision tree, Naïve Bayes and Support Vector Machine etc.

We would try building different models using different techniques ,validate, compare and come up with the best model which gives the generic ratings for the products that serves people to take a decision on the product purchase.

Below shown is the flow of one of our methods of implementing the project.



**S.M.A.R.T. OBJECTIVE:**

In the next 3 months, our team would have used 10 extra hours per week to build a model which gives the generic product rating for minimum of 120 products.

**DATA SOURCES:**

Data has been pulled from different sources like twitter, youtube, flipkart, amazon, snapdeal.

|  |  |  |  |
| --- | --- | --- | --- |
| **SOURCES OF DATA** | **SCRAPING TECHNIQUE** | **LIMITATIONS** | **METHODS USED TO OVERCOME LIMITATIONS** |
| E-Commerce sites | Web Scraper browser extension —> R script |  |  |
| Social Media (Twitter, Youtube) | Inspect elements —> curl command —> python script | Multiple requests from the same IP Web scraping bots fetch data very fast, which a human cannot.website became unresponsive. | Rotating IPs using VPN. Added sleep calls and some delays after crawling |

**Other limitations faced during data scraping**

Avoided to scrape websites that had a login as we got blocked easily. Without login, scraping has been performed.

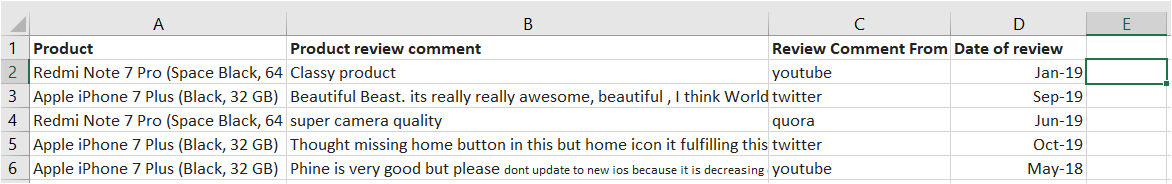
**PRELIMINARY DATA DESCRIPTION:**

Data set includes 4 variables. They are,

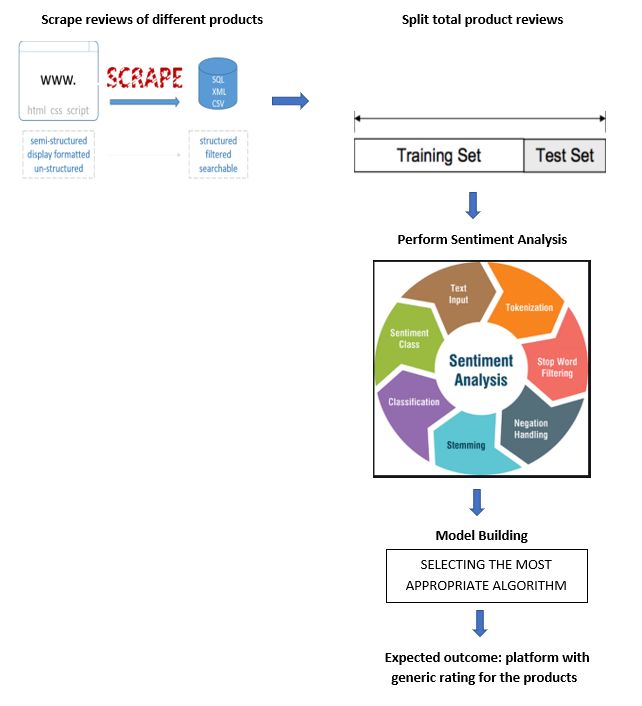
1. Product Name
2. Product review comment
3. Review Comment From
4. Date of review

**Dependant variable (output) :** Generic Rating of the product

**Sample of dataset:**



**FLOW DIAGRAM OF PROJECT**

****

**RECOMMENDATIONS EXPECTED OUT OF THIS PROJECT:**

Our recommendation on building the consolidated review platform through sentiment text analysis would in turn be the output which is being generated to the consumers through the model. The output here will provide an upper hand to the consumer in judging a product based on one's preference and make decisions.

There are few precautionary measures which are taken into consideration while building this model.

Since the basic step for data preparation is through Web Scraping of E-commerce and social media websites, there are limitations on the data that can be pulled out. For example, In Twitter, retrieval of data older than 7 days would be infeasible.

From a marketing standpoint, when a new product is released, the reviews given by users in the first few days of its launch would vary so much when compared to the reviews posted after a couple of weeks. Few factors influencing these reviews at an initial stage would include the look and feel, basic performance of the product and such more.

Thereby, the time we pull the data respective of product launch would have a minor influence on the reviews collected.

**References and Bibliography:**

1.<https://www.researchgate.net/publication/326200716_Sentiment_Analysis_Based_Product_Rating_Using_Textual_Reviews>

2.<https://www.researchgate.net/publication/330182227_SENTIMENTAL_ANALYSIS_OF_TWITTER'S_TOURISM_DOMAIN_DATA_USING_FEATURE_BASED_APPROACH>

3.<https://www.researchgate.net/publication/326482750_ALTAS_An_Intelligent_Text_Analysis_System_Based_on_Knowledge_Graphs_Second_International_Joint_Conference_APWeb-WAIM_2018_Macau_China_July_23-25_2018_Proceedings_Part_I>