**SYNOPSIS**

INTRODUCTION

The objective of this project is to extract sentiment from millions of reviews and analyze the impact they have on customer as well as business organization. The data set we used in our project is called Amazon product data collected from kaggle. Amazon is one of the largest and most known online vendor in the World of Ecommerce. People often gaze over the products and reviews (positive or negative) of the product before buying the product on amazon. The reviews on amazon are not necessarily of products but a mixture of product of product review and service review (amazon related or Product Company related). Consumers and sellers spend a large amount of time reading through ample amount of reviews of the product they are interested in, to find out what is perceived as good and bad about a product. Amazon currently has a feature that lets users filter reviews by popular keywords, which is still tedious and time-consuming for customers. The users have to read through numerous reviews to get the relevant information about the products that they need. Our model adds an additional layer on amazon reviews system, which extracts key aspects of a product, groups them, and determines their polarity.

PROBLEM DEFINATION

The aim of our model is to identify different aspects of a product that have been reviewed, and determine the polarity of the most common aspects. To ascertain which aspects to consider, we extract adjective-adverb pairs from the review text in Amazon Customer Reviews Dataset. The adverb in these pairs are used to improve the performance of sentiment analysis and the adjectives are used to determine polarity. Results from the model will be displayed, which shows most popular aspects of a product, grouped into positive and negative.

LITERATURE REVIEW

All Information in the world can be broadly classified into mainly two categories, facts and opinions. Facts are objective statements about entities and worldly events. On the other hand, opinions are subjective statements that reflect people’s sentiments or perceptions about the entities and events. Maximum amount of existing research on text and information processing is focused on data mining and getting the factual information from the text or information on the internet. Before we had the era of World Wide Web we were lacking a collection of opinion data, if an individual needs to make a decision, he/she needs to ask for opinions from friends and families manually. When a business organization needs to find opinions of the general public about its products and services, it conducted surveys on focused groups. But after the growth of Web World, especially with the drastic growth of the user generated content on the Web, the world has changed and so has the methods of gaining one’s opinion. One can post reviews of products at sites and express views on almost anything in forums, discussion groups, and blogs, which are collectively called the user generated content. As the emerging technology of connectivity grew so as the ways of interpreting and processing of user’s opinion information has changed. Some of the machine learning techniques like Naïve Bayes, Maximum Entropy and Support Vector Machines.

PROPOSED STUDY

The aim behind our model is that the aspects extracted from a set of reviews of a product  
can be similar or related to one other. So considering the most common aspects verbatim(adverb+adjective) for sentiment analysis might not be an accurate representation of the opinions in the reviews. Users may discuss the same features of a product in different words; clustering the aspects before determining the most frequently mentioned aspects would prevent the omission of key aspects. Additionally, it will also ensure that there is no redundancy. Once the aspects are clustered, the polarity for each cluster is calculated as the mean of the polarities of all aspects that belong to the cluster.

RESEARCH METHODOLOGY

The service and the product review’s polarity is the feedback;the user provides for the review. Finally, when a feature sentiment is extracted the sentiment phrase is sent to a polarizer method, this method basically returns +1 if the phrase is a positive sentiment else -1 if the phrase is a negative sentiment. Firstly, the phrases are tested for indirect opinions such as “REDMI NOTE7 is no better than REDMI NOTE7 PRO”, the test phrase is tested for certain pre-defined phrases that were found during manual analysis of reviews. Next, if the phrase test fails, the review is tested for the word “not” if the word not exists then everything after not is polarized meaning every word after not is tested for whether it is a positive word or a negative word and consecutive words polarity are added and finally negated, for example “Phone is not good” this phrase is classified as negative as the word “good” is negated by the word “not”. Lastly if “phrase” and “not” test fail the test phrase is broken down into words and polarity of each word is found from a dictionary of sentiment words bifurcated as good and bad words and collective polarity is considered i.e. if the sum is below 0 the outcome is negative (-1) else outcome is positive (+1).

Rules for feature extraction The following are some rules that our model uses to extract feature and its sentiment:

1. Adjective + Noun

2. Noun +Adjective

3. Adverb + Noun

4. Adverb +Adjective + Noun

5. Noun + Adverb + Verb

6. Noun + Verb

7. Noun + Verb

8. Noun +Verb + Noun

9. Noun +Determiner + Adjective

10. Noun + Verb + Adverb

11. Noun + (verb or Adjective or Adverb)

WORK PLAN

First, we need to do Natural Language Processing in order to perform positive and negative analysis for word counting. Sentiment Analysis is one of the most popular applications of NLP.To perform sentiment analysis we require the tool which is python 3.0. (spyder). We built APIs for data extraction and create file path for the data.

STEPS REQUIRED TO PERFORM:

1.Create a new python file and import the following Packages.

2.Define a functions to extract features.

3.We need training data for this so will use Amazon product Review.

4.Extract the feature.

5.Cleaning of feature which is being extracted.

6.Analyze the sentiment of a piece of text with SDK.

7.Lets separate it into positive and negative review.

8.Divide the data into training and testing datasets.

9. We will also preprocess the text data before, in order to extract better feature from clean data.

10.Use panda and matplotlib for visualization of the result of the work.

STARTING WITH SOME BASIC TECHNIQUE OF NATURAL LANGUAGE PROCESSING TO THE ADVANCEMENT OF THE NATURAL LANGUAGE PROCESSING:

We can use text data to extract a number of features even if we don’t have sufficient knowledge of  [Natural Language Processing](http://courses.analyticsvidhya.com/courses/natural-language-processing-nlp?utm_source=blog&utm_medium=UltimateGuideTextDataarticle).

* First step of NLP- Text Processing
* Extracting Named Entities from Text
* Feature Engineering for Text
* Mastering the art of Text Cleaning
* Social Media Information Extraction
* Interpreting patterns from text

**WEB SCRAPING**

* Scrape product details that you can’t get with the product advertising API.
* Monitor products for change in price, Stock count/Availability, Rating, etc.
* Analyze how a particular Brand sells on Amazon
* Find customer Opinion from Amazon Product reviews
* We will build a scrapper that can go to any amazon product page using an ASIN (a unique ID Amazon uses to keep track of products in its database)

For example : <https://www.amazon.com/Samsung-A10-Infinity-V-Smartphone-International/dp/B07Q6ZNJNT/ref=sr_1_3?crid=11OGCSA4NZQAU&keywords=samsung+mobile+phone&qid=1559632312&s=gateway&sprefix=samsung+mobile+%2Caps%2C1269&sr=8-3>, the ASIN in B07Q6ZNJNT.

* The next is to builds a script that goes to each one of those product pages, downloads its HTML and extracts the fields you need – e.g. Product Title, Price, Description, etc. Once we extract this information and save it to a JSON file. Since, we already have the list of products.

PROPOSED CONTENT OF THESIS

TOOLS AND TECHNIQUES

The sentiment analysis is a complex with sequence of process that involves 5 different steps to analyze sentiment data. These steps are:

• Data collection: The first step of sentiment analysis begins with the collection of data from user generated content contained in blogs, forums, social networks. These data are disorganized, expressed in different ways by using different vocabularies, slangs, context of writing etc. Manual analysis is almost impossible for the millions of data on may be millions of languages. Therefore, text analytics and natural language processing are used to extract and classify;

• Text preparation: This consists in cleaning the extracted data before analysis. Non-textual contents and contents that are irrelevant for the analysis are identified and eliminated;

• Sentiment detection: The extracted sentences of the reviews and opinions are examined. Sentences with subjective expressions (opinions, beliefs and views) are retained and sentences with objective communication (facts, factual information) are discarded;

• Sentiment classification: In this step, subjective sentences are classified in positive, negative, good, bad; like, dislike, but classification can be made by using multiple points;

• Presentation of output: The main objective of sentiment analysis is to convert unstructured text into meaningful information. When the analysis is finished, the text results are displayed on graphs like pie chart, bar chart and line graphs. Also time can be analyzed and can be graphically displayed constructing a sentiment time line with the chosen value (frequency, percentages, and averages) over time. The paper provides an overview of studies on the sentiment.

Supervised learning is a widely used solution for classification purpose and is been used in most of the sentiment classification techniques. Techniques for sentiment classification include SVM, Neural Network and Decision Tree Classifiers. Other commonly used algorithms include K-Nearest Neighbour, Bayesian Network. The Naïve Bayes Classifier is one of the most commonly used classifier for text data. Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes model is easy to build and useful for very large data sets. The method calculates prior probabilities from training data and computes the posterior probability of the document based on the prior probability values. Maximum entropy classifier is a probabilistic classifier which converts labelled feature sets into vectors by using encoding. Support vector machine is non probabilistic approach which is used to separate data linearly and nonlinearly. It determines the separators in search space which can best separate the classes. SVM is suited for text data and is used in most of the sentiment analysis techniques.