Research Project Proposal

Sentimental Analysis of Amazon Dataset

Name: Ajinkya Kunjir

NEU ID: 001869353

Date: February 2, 2018

**Abstract**

Sentiment analysis is the measurement of positive and negative language. It is a way to evaluate written or spoken language to determine if the expression is favorable, unfavorable, or neutral, and to what degree. Today’s algorithm-based sentiment analysis tools can handle huge volumes of customer feedback consistently and accurately. Paired with [text analytics](https://www.clarabridge.com/text-analytics/), sentiment analysis reveals the customer’s opinion about topics ranging from your products and services to your location, your advertisements, or even your competitors.

With our dataset of ‘Amazon’s Fine Food Reviews’, we aim to build a prediction model where we will be able to predict if a review is positive or negative. In addition to that, we shall also filter out the products which have got the most reviews and ratings. We will be focusing on the textual reviews along with the food ratings to create accurate prediction. Often ratings are ignored in the interest of textual reviews which might lead to improper and irregular reviews. We aim to take both in account and hence counter the problem.

**Background**

With online shopping on rise, electronics and clothing seem to have taken the backseat, and grocery has been the major driver for the online vendors. With the youth becoming increasingly dependent on the internet, the idea of shopping for groceries and food items has become highly relevant and the online shopping vendors have started tapping into the potential of the same. Online vendors can provide items not available in a region at almost no difference in prices and thereby making it easier for the consumer to get access to what they want easily.

For this, online reviews and ratings of a product has become very important for analysis and evaluation of the quality of the product. The need to filter out the ratings and reviews according to the required needs will prove out to be beneficial not only to the company, but also to the consumer. With internet becoming the new way to raise your voice, reviews and ratings play a significant role.

**Algorithms**

The major algorithm primarily being used for sentimental analysis are:

#### Machine Learning for Sentiment Analysis:

1. Supervised Learning
   * Naive Bayes
   * Support Vector Machine (SVM)
2. Unsupervised Learning
   * Dictionary-based approach
   * Corpus based approach
   * Clustering or Scoring algorithms

In our project we will primarily focus on Naïve Bayes and the Dictionary-based approach. We will compare the results of the same at the end of the project. We shall tweak the algorithms according to the requirement of the project to get accurate results.

**Data Sources**

Our primary source of data is Kaggle. We shall be using the dataset available on Kaggle.com.

The link for the dataset is: <https://www.kaggle.com/snap/amazon-fine-food-reviews/data>

The data is in .sqlite format. The data is made up of ~500,000 rows and 10 columns.

**References**

[1] <https://www.kaggle.com/snap/amazon-fine-food-reviews/data>

[2] <https://en.wikipedia.org/wiki/Naive_Bayes_classifier>

[3] <http://www.ijeert.org/pdf/v3-i1/9.pdf>

[4] <https://lct-master.org/files/MullenSentimentCourseSlides.pdf>