Project Proposal

Problem Statement

The Task here is to do sentiment analysis and classify customer sentiment as either positive or negative based on review text. We also predict the customer overall rating (1 being the minimum and 5 maximum) for cellphones based on customer reviews.

The data for this project is from Amazon’s customer reviews on cellphones. This dataset focuses on both unlocked and locked carriers, and scoped on ten brands: ASUS, Apple, Google, HUAWEI, Motorola, Nokia, OnePlus, Samsung, Sony, and Xiaomi.

Who cares?

Sentiment Analysis can be used by any businesses that care about the opinion of their customers. In our scenario, Amazon is the client who wants to analyze sentiment of their customers based on the reviews for various cellphone brands obtained on their website.

Sentiment analysis (or opinion mining) is a natural language processing technique used to determine whether data is positive, negative or neutral.

Benefits of Sentiment Analysis: -

1. Helps improve customer service - Sentiment analysis provides businesses an idea of the customer opinion about their services. It helps identify problems and persuades the organization to look for a solution. The management make use of sentiment analysis to improve their customer services.
2. Helps develop quality products- The responses from the customers can be used as a guideline to improve the service quality, better future product development, reduce customer churn or improve how the product is presented.
3. Develop new marketing strategies- With more data and information gathered through sentiment analysis, the organizations could develop an effective marketing strategy. By observing the customers’ conversations on their social media and detect the specific key messages related to your brand, specific marketing campaigns can be designed for the target consumers.
4. Improves crisis management- Frequent monitoring of the customers’ responses or opinions towards a brand would help to identify any issues quickly.

Source of Data

The data was collected from Amazon’s online shopping website on Dec 26,2019. The data contains 2 csv files – items.csv and reviews.csv.

items.csv contains retrieved (read: scraped) items from Amazon.com search results using generated URL and specific query string to search only specific brands and has minimal 1 star review.

reviews.csv contains reviews for previously retrieved items at items.csv but not with columns from items.csv.

What approach do we use to solve the problem?

To implement Sentiment Analysis, we need the follow these steps: -

1. Data retrieval with web scraping
2. Data Cleaning/Text pre-processing
3. Exploring the data
4. Modeling
5. Sentiment Analysis

Data retrieval – If data is not readily available in files, we will have to do web scraping using libraries such as BeautifulSoup or Scrappy to pull data out of HTML, XML or other Markups.

Data Cleaning is the process of cleaning and unifying messy and complex data sets for easy access

and analysis.

Text preprocessing -Almost every [Natural Language Processing (NLP)](https://en.wikipedia.org/wiki/Natural_language_processing) task requires text to be preprocessed before training a model. Text preprocessing commonly involves removing punctuation, converting to lowercase, tokenization, stemming, stop word filtering, parts of speech tagger etc.

Modeling – To predict ratings based on reviews we could use regression models, or we could even try using Naïve Bayes classifier.

To classify customer sentiments there are 2 approaches –

1.Supervised ML approaches or Deep learning approaches

2. Unsupervised Lexicon based approaches.

For the first approach we typically need pre-labeled data. Since we can’t obtain that, we could focus on the second approach. Various popular lexicons are used for sentiment analysis, including the following:-

AFINN lexicon

Bing Liu’s lexicon

SentoWordNet

MPQA subjectivity lexicon

VADER lexicon

TextBlob lexicon

Among these, VADER is said to be more suitable for online data.