**Voice of the Customers: Mining Online Customer Reviews for**

**Product Feature-based Ranking**

**Vivek Gupta (G39431539)**

**Omkar Dixit (G42396299)**

George Washington University

Project

**April 21, 2016**

**Supervised By: Dr. Abdelghani Bellaachia**

**ABSTRACT**

This project is about finding out sentiments about any product. Twitter data was used for analysis. Iphone product was selected for analysis. Sentiment analysis was done according to location and features such as Screen, price, battery of a phone. Python, nlp tool kit was used for sentiment analysis. For visualizing results, WEKA was used.

**Table of Contents**

1. **Introduction 4**

**1.1 Definition of problem 4**

1. **Approach 4**

**2.1 Data collection 5**

**2.2 Data pre-processing 6**

**2.3 Knowledge discovery 7**

1. **Results and discussion 7**
2. **Limitations and future work 8**
3. **Reference 9**

**1 Introduction and motivation**

Now a days, most of the people are shopping online and customers are becoming smarter. Before purchasing anything, they check feedback on Google. And many people write reviews online on product web site or social networking websites such as Twitter. Thus, it is very important for company to know feedback about product so that they can bring new product with reduction in features which are not so popular.

Let’s take an example of star bucks. They launched a drink with 10$price. And checked reviews online and they found that price is too high. Thus they reduced price within hours and thus, product was successful.

Thus, it is very important for company to find out positive and negative points about products.

**1.1 Definition of the problem**

Analyze sentiments for product iphone from twitter based on location and features like Screen, Battery.

**2 Approach**

As twitter is becoming very good place to get reviews, we used to get data. Data mining processes needs to be applied to twitter data. Overall process can be summarised as shown in figure below

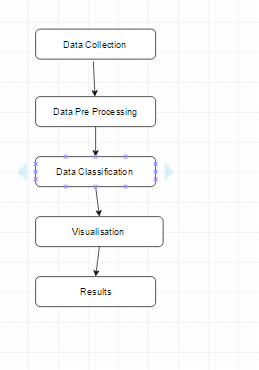


Figure 2.1( Flowchart of process)

Let’s, look at each process in detail.

**2.1 Data Collection**

Data was collected from twitter developer api.

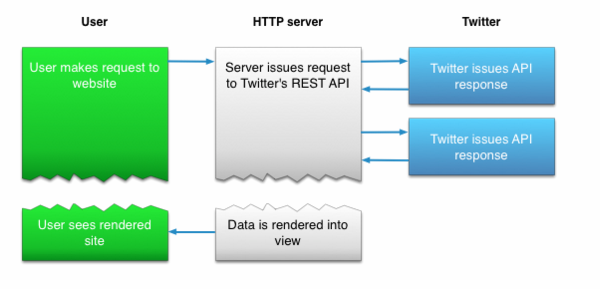


Figure 2.2 Twitter API (<https://dev.twitter.com/overview/documentation>)

**2.2 Data Pre processing**

Twitter data from twitter api cannot be directly used for sentiment analysis. It contains many junk characters, urls which are not required for sentiment analysis. Also tweet data contains text from other languages also. Thus, twitter text pre processing is required. First data filtering was carried out. Filtering was done based on location and language. English language tweets were selected.

After filtering data cleansing was carried out. Setps for data cleansing:

1. Removed ascii values.

2. Converted all tweet text to lowercase.

3. Removed URLs as they won’t play any role in finding sentiment.

4. Still some Thai and Latin tweets were not removed Thus, they were removed using regular patterns.

5. Split the tweets according to ‘,’ ‘;’’.’’and’’but’ etc.

6. Removed stop words from sentences.

7. Data was again filtered according to features like battery, screen, and price.

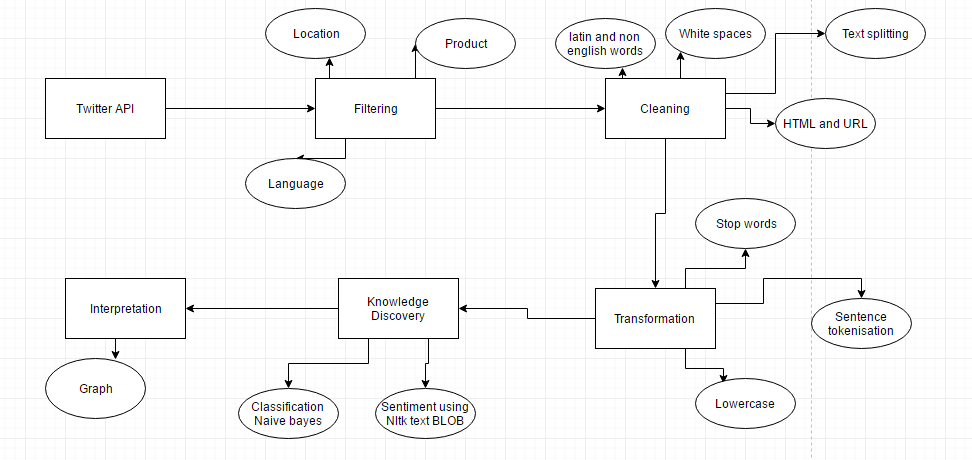


Figure 2.4 Overall flowchart of process

**2.3 Knowledge discovery from data**

As data was converted according to features, we can use sentiment analysis. For sentiment analysis we used 2 approaches:

1. Use word corpus to find out sentiment for sample data iteratively run Naïve Bays’ algorithm.

2. Use Part of Speech tagging for finding adjectives for features and according to features and then use word-corpus to categorize sentiments.

Overall flowchart for knowledge discovery is as follows

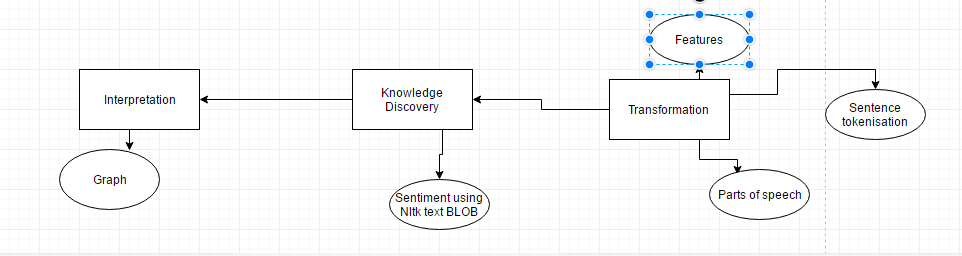
****

Figure 2.5 Classification process

**3 Tools Used**

Python: For twitter data pre processing.

NLP Toolkit: For sentiment analysis

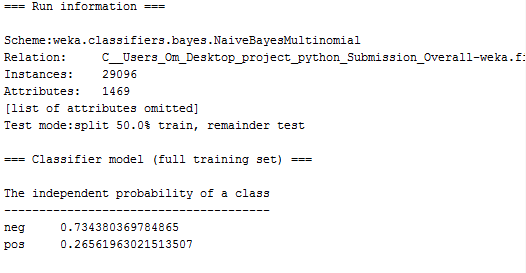
Weka: For applying algorithms and visualisation of results.

**4. Results**

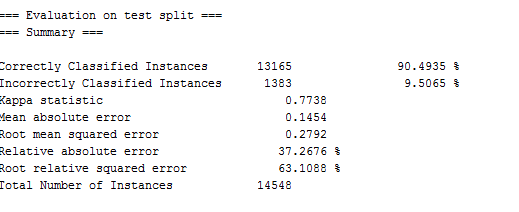
Let’s look at result. Procedure to run is described in readme file and description of each program is written in project manual.

**4.1 Sentiments for price**

As price is important feature of a product, twitter comments were filtered according to price. Adjectives were extracted from data and according to polarity of adjectives, comments were classified as positive and negative. We used NaiveBays’ Multinomial classifier to extract results. We used 50% training dataset and 50% test dataset.



Accuracy was 90%.



Results were plotted on Weka. Results are shown in figure 4.1

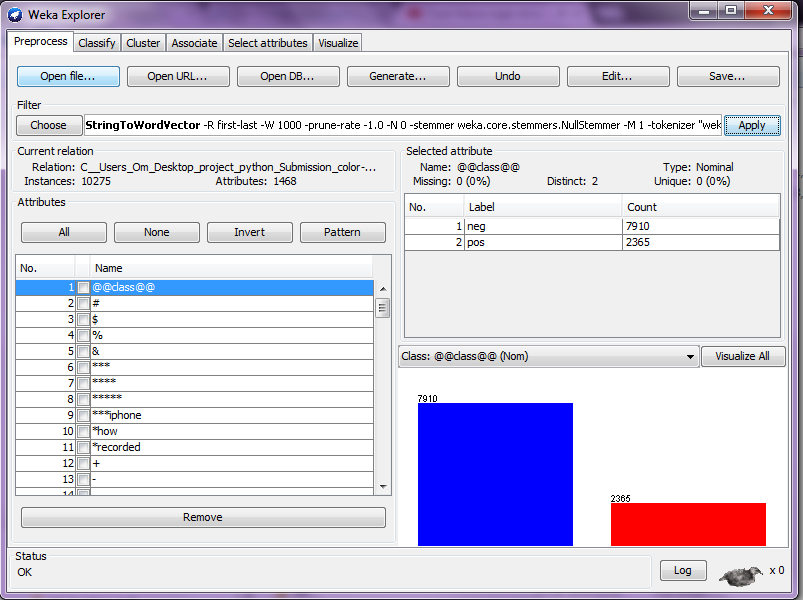


Figure 4.1 Sentiment analysis based on price

**4.2 Overall sentiment in India**

It is important to find overall sentiment based on location.Adjectives were extracted from data and according to polarity of adjectives, comments were classified as positive and negative. Results were plotted on Weka. Results are shown in figure 4.2

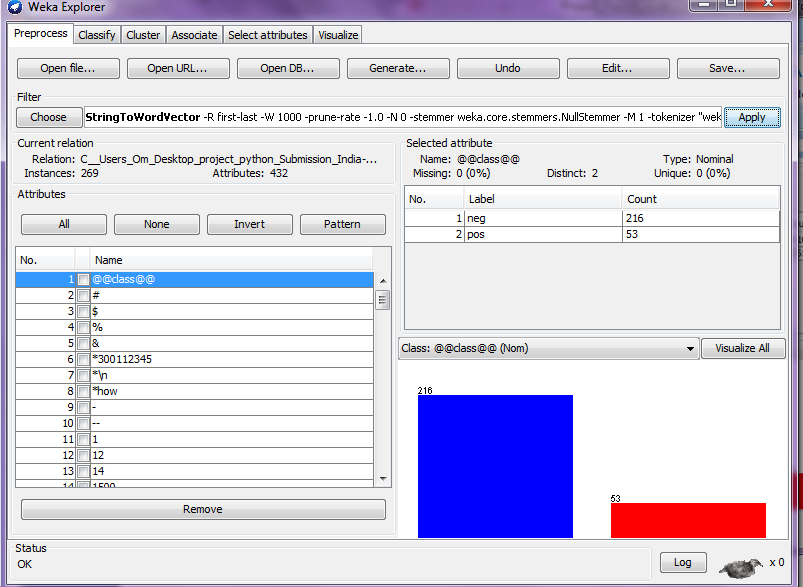


Figure 4.1 Sentiment analysis for Indian tweets

**4.3 Sentiments in battery in NY**

Tweets were extracted for New York location and for battery. Results are plotted in Weka

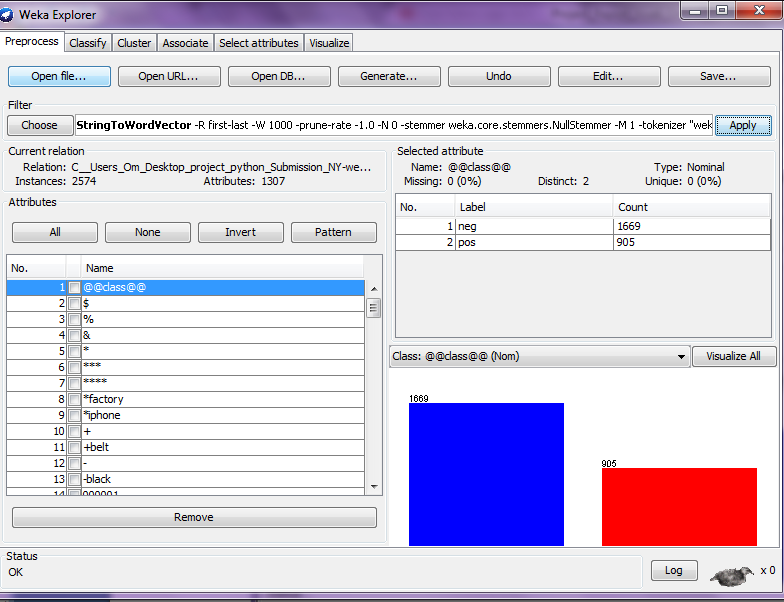


Figure 4.3 Sentiment analysis for camera for New York

**4.4 Overall sentiments**

Overall sentiments were classified and results were plotted in WEKA.

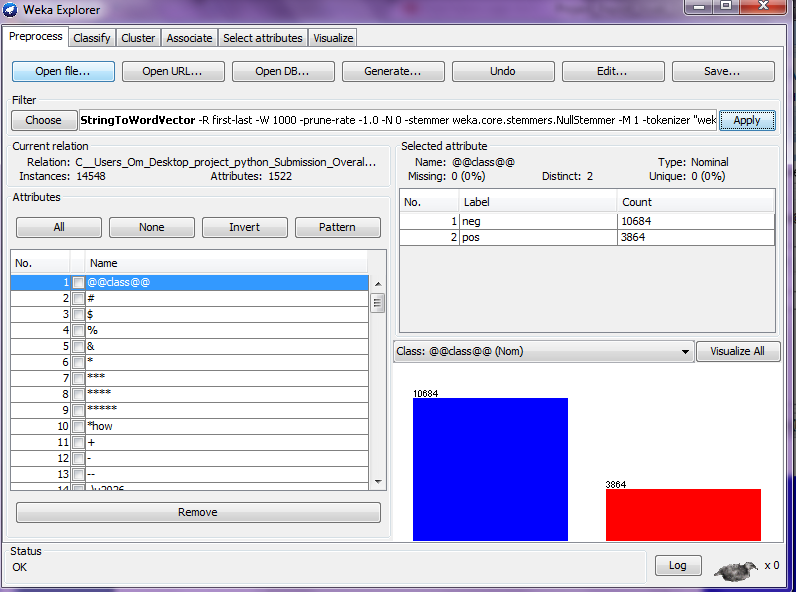
****

Figure 4.4 Overall sentiment analysis for all English Tweets

According to results we can conclude that,

Iphone is not popular because of it’s price in India and battery in New York.

Overall iphone is getting negative buzz in Twitter.

**Limitations and Future work**

Sarcastic analysis was not implemented. It can be implemented further.

Detection of attributes of product according to LDA can be implemented further. And it can be generalised for any product.

**References**

1. <https://dev.twitter.com/overview/documentation>

2. <http://nlp.stanford.edu/>

3. Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." Proceedings of the workshop on languages in social media. Association for Computational Linguistics, 2011.

4. Liu, Bing. "Sentiment analysis and opinion mining." Synthesis lectures on human language technologies 5.1 (2012): 1-167.