BIA- 660 Web Analytics

Midterm Report – (Project Team 4)

Members:

Madhura Milind Nagaonkar

Pratibha Dwivedi

Preeti Negi

Shashank Mysore Bhagwan

Motivation:

For ages, we have been witnessing a tough war between Apple and Samsung with respect to iPhone and Galaxy smartphones. But in the recent past, we saw development in the field of Surfaces. Microsoft who is a pioneer in developing Surfaces, went all out in terms of features, with their Surface Pro Tablet. To counter this, Apple also has launched their version of Surface, called 'iPad Pro' with smart keyboard attachment. With Apple and Microsoft already in the Surface race, how can Samsung be left behind. In 2016, they launched Galaxy Tab Pro. Surfaces are a flexible, lightweight and stylish alternative to the traditional laptops. Our main aim is to guide customers when they want to buy these products, by providing them the information of the features they care about.

Methodology:

- We scrapped the data i.e. customer reviews from best buy and amazon website. We used selenium webdriver as we need to automate the web browser to access the information.
- We saved the raw data with fields as rating and review in a csv file and used that for further process. The csv we collected were named as ipadpro_raw.csv, surfacepro_raw.csv and samsungpro_raw.csv.
- After collecting the csv files for the three products, we used natural language processing techniques to clean the data. We removed punctuations, stopwords and converted it to lower case. We saved the generated data in clean csv files which are used for further analysis. The csv's were named as ipad_pro_clean.csv, surface_pro_clean.csv and samsungpro_clean.csv.

Data Sources:

1) Best buy website

Ipad Pro	**link-** https://www.bestbuy.com/site/reviews/apple-10-5-inch-ipad-pro-latest-model-with-wi-fi-64gb-space-gray/9093027?page=
Surface Pro	**link1-** https://www.bestbuy.com/site/reviews/microsoft-surface-pro-12-3-touch-screen-intel-core-m-128gb-with-black-type-cover-latest-model-platinum/6111206?page= **link2-** https://www.bestbuy.com/site/reviews/microsoft-surface-pro--12-3--intel-core-i5--8gb-memory--256gb-solid-state-drive-latest-model-silver/5855921?page= **link3-** https://www.bestbuy.com/site/reviews/microsoft-surface-pro--12-3--intel-core-i5--4gb-memory--128gb-solid-state-drive-latest-model-silver/5855919?page=

2) Amazon website

	link1- https://www.amazon.com/Samsung-Galaxy-TabPro-Performance-TouchScreen/product-reviews/B06XB8FFG8/ref=cm_cr_arp_d_viewopt_srt?ie=UTF8&reviewerType=all_reviews&sortBy=recent&pageNumber= **link2-** https://www.amazon.com/Samsung-Galaxy-Windows-Silver-SM-W720NZKBXAR/product-

Samsung Tab Pro	reviews/B06XF2LCMT/ref=cm_cr_arp_d_viewopt_srt?ie=UTF8&reviewerType=all_reviews&sortBy=recent&pageNumber= **link3-** https://www.amazon.com/Samsung-Wifi-Tablet-Black-SM-W700NZKAXAR/product-reviews/B01AZ7LS94/ref=cm_cr_arp_d_viewopt_srt?ie=UTF8&reviewerType=all_reviews&sortBy=recent&pageNumber=

Analysis:

As per the methodology followed for scraping and NLP, we have gained valuable output in each scraped csv. Each scraped csv represents a separate product. The scraped data was then pre-processed(cleaned) and used as input to the code for sentiment_evaluation.py to perform sentiment analysis on positive and negative words from the reviews to evaluate overall sentiment of the reviews. Thus, leading to sentiment evaluation or calculation of overall percentage of positive and negative reviews for a product.

Moving forward, to implement accuracy functionality/performance evaluation, we are using the same cleaned data as input to the code of sentiment_performance_evaluation.py. To obtain a respectable accuracy, we assigned the Ratings of '5', '4', '3' with positive sentiment and Ratings of '1', '2' with negative sentiment. These sentiment values are compared to the sentiment analysis function developed using the positive and negative words text files. Thus, evaluating performance of our sentiment analysis, by comparing the sentiments based on the customer ratings and the sentiments calculated from the customer reviews

As mentioned in the Motivation section, Ipad Pro and Surface Pro are two legendary devices and hence the reviews for these products are more in number. For samsung Tab pro we could acquire fewer reviews since it was released in 2016. The reviews for Samsung Tab Pro also consist of reviews for Samsung Galaxy Book which is the next version of Tab Pro released in 2017.

We have performed sentiment analysis of the reviews. To get a better understanding, we have merged the output of all products and displayed in a table as follows-

COMPARISON METRICS	Ipad Pro	Surface Pro	Samsung Tab Pro
Number of Positive Reviews:	2051	1093	213
Number of Negative Reviews:	283	168	84
Percentage of Positive Reviews:	87.87%	86.68%	71.72%
Percentage of Negative Reviews:	12.13%	13.32%	28.28%
Accuracy:	91.563%	90.065%	83.893%

Analyzing the above table, it is very clear that the Ipad Pro and Surface Pro have better customer ratings with respect to the newly released Samsung Tab Pro. Considering the popularity of apple in United States, we can see the difference in the number of customer reviews on Best buy website for Surface Pro and Ipad Pro.

The Accuracy was calculated based on the logic that ratings of 5,4 and 3 will be positive and ratings of 1 and 2 will be negative. This was then compared to the actual logic of sentiment analysis. We are receiving an accuracy of almost 91 percent for Ipad and Surface, and an 83 percent accuracy for Samsung. This goes to show that, the customers have rated to product appropriately as per their description in the review.

Next Stage:

- As the very first step, we will select the reviews randomly and will assign labels to each review.
- We will conduct supervised training by using Convolutional Neural Network model which is a multi-label classification technique used to classify reviews into designated groups. This will help us to scrap the customer reviews and find valuable information.
- Then, for each feature our group will conduct sentiment analysis to know about the customer opinions. We will use positive and negative dictionary, count positive and negative words associated with each feature of the tablets.