

SALES OPPORTUNITY

CSC 591 Data-Driven Decision Making

Team 3 - Members

Vivek Gopalakrishnan(vgopala2)

Anbarasi Manoharan(amanoha2)

Moharnab Saikia(msaikia)

Neela Niranjani(nvengat)

Rupaj Soni(rosoni)

Guide

Michael Kowolenko

NC STATE UNIVERSITY

Table of Contents

1. INTRODUCTION	1
1.1 Background	1
1.1.1 Opinion mining	2
1.1.2 Product Quality Improvement	2
1.1.3 Competitive intelligence	3
2. PROBLEM STATEMENT	3
Generic Question	4
Project Question	4
2.1 Subproblems	4
2.1.1 Prioritization	4
2.2 Assumptions	5
2.3 Bias	5
3. METHODOLOGY	5
3.1 Data	5
3.2 Tools	8
3.3 Feature Extraction and Sentiment Analysis	9
3.4 Truth Tables	10
3.5 Query Processing	12
4. OBSERVATIONS	13
5. CONCLUSION	16
6. FUTURE WORK	16
7. REFERENCES	16
8. APPENDIX	17

ABSTRACT

The increasing pervasiveness of the Internet and the proliferation of shopping websites, blogs, review websites, forums and social networks presents a new set of challenges and opportunities in the way information is searched and retrieved. Even though facts and research data play an important role in developing a product, opinions have become increasingly important as they provide the company with a permanent lighting to its competitive environment. Thus, there opens up an opportunity where Enterprises can mine such information to determine how users perceive their products and how they stand with respect to competition. This information introduces a new source that can also help the company to identify, analyze and manage the various risks associated with its products. In our report we explain an application we developed which makes recommendations on which features to include in the next iteration of the iPhone by analyzing the sentiments/opinions of the users on its current features and well as the features of its competitors.

1. INTRODUCTION

1.1 Background

In the present time, people are becoming increasingly involved in sharing their personal opinions on the internet through different sources like online shopping sites, blogs, review websites etc. These opinions can help in understanding the limitations of a product and aid in early detection of defects. This would simplify risk management, reducing future liabilities, as well as help the company executives to make sound business strategies.

Due to the ease of expressing views on the internet and its wide proliferation, there has been a rapid increasing in the quantity of user generated content, such as user feedback reviews, blogs, online forums, discussion groups, social media etc. Different from other kinds of online textual information, user generated contents are people-centric and contain a lot of subjective information. Extracting these subjective texts and mining user opinions in the text is valuable to both customers and manufacturer. For customer, one could search the

opinions to find the opinion of existing users. For manufacturer, apart from using traditional way such as customer surveys, they can gather customer reviews from these online sources to make product improvements by understanding the preferences of the customers.

1.1.1 Opinion mining

Currently, review websites, blogs and social media play a very important role in different sectors and different businesses. For example, for a company that wants to know the impact of its products in comparison to its competitors may exploit customers opinions. Due to the vastness in volume the ability to analyze a set of online reviews and produce an easy to digest summary is a major challenge and is very difficult to achieve using traditional techniques. This is where opinion mining comes into the picture. Opinion Mining or Sentiment Analysis, a subtask of text mining, is a process that analyses the conversations around an event, a topic or a product, based on a system that automates this process. As the major components of opinion mining, we can mention subjectivity analysis, affect analysis, emotion analysis, contextual polarity (positive or negative) and the polarity strength (weakly positive, mildly positive, strongly positive, etc.) of a document or comment. In this project, we focussed on product opinion mining where the polarity of the opinion concerning a product's feature or characteristic was examined.

1.1.2 Product Quality Improvement

Product quality improvement is a formal approach to the analysis of performance of products and systematic efforts to improve it. It is a form of ongoing effort to make performance better. Product quality improvement aims at ways to improve customer satisfaction by meeting their needs. In the mobile industry quality efforts focus on improving device performance, reducing cost, reducing customer complaints, reducing field returns, improving employee satisfaction level, improving supplier performance, and these all combined aim at increasing the market share. Product quality improvement is of prime focus at any industry especially the mobile industry as declining quality directly impacts the confidence of the customers on the company. This results in them turning to a competing device. Also once lost, its very difficult to gain the previous level of

value in the market as this is a very dynamic market with stiff competition. For example, in the last few months the issue of samsung phones exploding has brought bad reputation to the company and subsequently decreased their share values to one of the lowest in last few years. To revive from this they will need some significant improvement as well as unique feature in the next iteration of their flagship phones.

1.1.3 Competitive intelligence

According to Bartes, Competitive Intelligence seeks to predict the future, and the strategic company decisions based on these predictions. Lubica defined competitive intelligence as the process of monitoring the competitive environment and the competitors, in which, information gathering, analysis and distribution of the obtained results, is carried out gradually so that they can support the efficient business activity and its ability to make qualified decisions, especially in relation to its competitors.

Companies must be able to develop new knowledge about its competitors in an increasingly complex and fast-moving economy to maintain levels of innovation and gain a competitive advantage. Therefore, the importance of Competitive Intelligence in companies practically becomes a necessity and widely accepted.

2. PROBLEM STATEMENT

It's the question; not the technology - Michael Kowolenko

Since the question is the main driving factor in the whole decision making process we devoted significant effort in determining the question before deciding on what technologies to use.

The initial topic of our project was “Sales Opportunity”. After much discussion we finally decided upon the following unambiguous question as our base for the project.

“Create a data-driven application which provides manufacturers with recommendations about what features to include in the next iteration of the iPhone to make it more appealing to users and thereby increase its sales.”

The main consideration in designing the application were:

- Provide a easy to use and intuitive UI
- Include the preferences of the user in the decision making process
- Makes sensible/valid recommendations
- Powerful visualizations
- Dynamic: ability to handle changes in data over time

To further increase the accuracy of the solution we considered the following questions to direct our decision making process.

Generic Question	Project Question
What do you want to know?	How to design a better iPhone 7 model
What problem or opportunity do you want to explore?	Analyze the drawbacks and reviews of iPhone 7 and augment the model to be more appealing to the users
What customer needs do you want to serve?	Present a phone model with features that match customer preferences and requirements
What capabilities do you want to test with the market?	What features, widely liked by customers do competitor products provide?
What new markets do you want to explore?	Compare iPhone 7 with other competitor products and analyze what features are missing in iPhone 7 that makes it less approved by users
What new sources of solutions do you want to explore?	Social media, blogs, expert reviews, customer ratings

2.1 Subproblems

After structuring the problem and coming up with our problem statement, the next step in the critical thinking process was to break the overall problem down into smaller, more manageable subproblems.

2.1.1 Prioritization

Prioritizing the subproblems is a key step to a successful application. There might

multiple problem that can solved in a domain. Addressing the sub problem that directs us most towards the answer has to be selected. It is an important step in the process of critical thinking. Sub problems were delegated to be dealt by different members of the team. However, some very solid subproblems general to the entire group consist of the weight-ordered recommendation model, truth table, infrastructure/tools, and measuring performance.

2.2 Assumptions

This project is aimed at proposing a model to improve product quality by means of analyzing customer opinions. The work has various limitations and assumptions such as: This project focuses on a mobile phone product, more specifically the iPhone 7/7 Plus . The next iteration of the product is taken into consideration to use. For the data the language chosen was English. We also assumed that the discussions made in the previous decade don't mean anything and we tried to filter data to the most recent years. We set a mark at 2014 to eliminate unwanted feature preferences which are already a core feature of the phones these days.

2.3 Bias

Choosing the top phones to perform our data extraction was our bias in the experiment. However, we made sure we included a fair sampling of the high-end phones. Also, our sentiment extractor comprises of both dynamic and static features. The static features to map dynamic features to a fixed set of clusters. This might introduce a bias in terms of having unproportionate clusters.

3. METHODOLOGY

3.1 Data

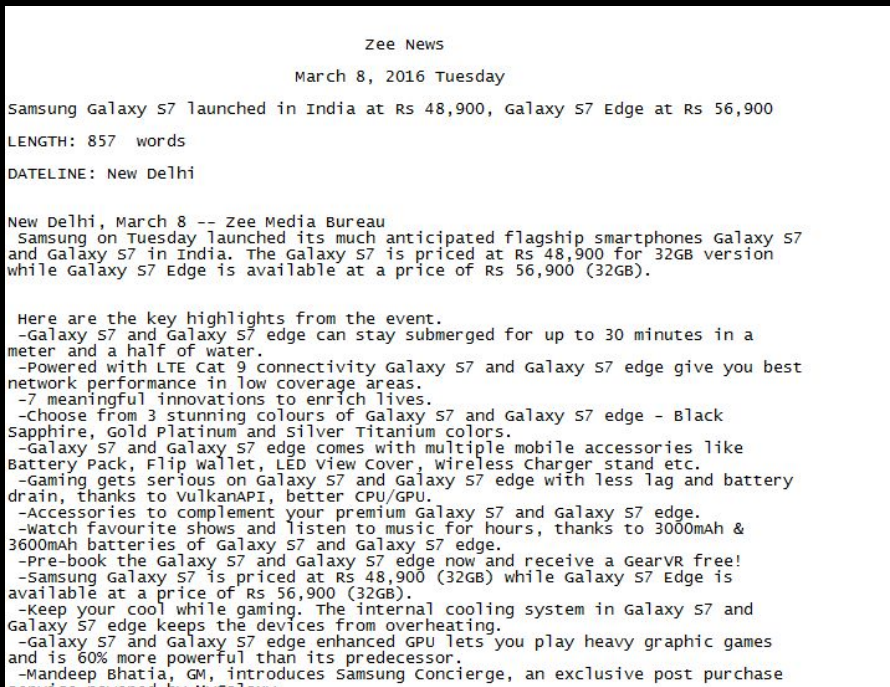
The data is collected in all forms- structured, semi-structured and unstructured.

Structured data is collected from the users/ manufacturers in the UI. It is done to give priority of one feature over another for their future iPhone. This results in a customizable phone for a customer/manufacture based on their personal choice. Priority data is clearly structured as the most important features were found and

given to the user and was asked to prioritize them from most wanted to least wanted. This creates a clear ranking that can be used in the recommendation system.

The unstructured data comes in form of tweets and reviews from blogs, websites. The reviews, articles, blogs, etc. data was obtained through the Lexis Nexis Academic Content Tool which collects data from all these sources and provides us in a single query. Around 1000 documents were collected for each of the phones used in the project.

The twitter data was collected through the Twitter Streaming API. We collected the twitter data using keyword match as well as hashtag filtering. For each of the phones, we had a set of variations of the phone name used for this purpose. Upon collecting the data, we filtered the relevant tweets. There was one interesting observation in this data. The tweets collected using the keyword match was more irrelevant to the mobile devices when compared with the tweets collected using the hashtags. Hence, we proceeded with only using the hashtag data for the feature extraction and sentiment analysis. This ensured that we obtained some specificity and context while processing the twitter data.



Zee News
March 8, 2016 Tuesday
Samsung Galaxy S7 launched in India at Rs 48,900, Galaxy S7 Edge at Rs 56,900
LENGTH: 857 words
DATELINE: New Delhi
New Delhi, March 8 -- Zee Media Bureau
Samsung on Tuesday launched its much anticipated flagship smartphones Galaxy S7 and Galaxy S7 in India. The Galaxy S7 is priced at Rs 48,900 for 32GB version while Galaxy S7 Edge is available at a price of Rs 56,900 (32GB).

Here are the key highlights from the event.
-Galaxy S7 and Galaxy S7 edge can stay submerged for up to 30 minutes in a meter and a half of water.
-Powered with LTE Cat 9 connectivity Galaxy S7 and Galaxy S7 edge give you best network performance in low coverage areas.
-7 meaningful innovations to enrich lives.
-Choose from 3 stunning colours of Galaxy S7 and Galaxy S7 edge - Black Sapphire, Gold Platinum and Silver Titanium colors.
-Galaxy S7 and Galaxy S7 edge comes with multiple mobile accessories like Battery Pack, Flip wallet, LED View Cover, Wireless Charger stand etc.
-Gaming gets serious on Galaxy S7 and Galaxy S7 edge with less lag and battery drain, thanks to VulkanAPI, better CPU/GPU.
-Accessories to complement your premium Galaxy S7 and Galaxy S7 edge.
-Watch favourite shows and listen to music for hours, thanks to 3000mAh & 3600mAh batteries of Galaxy S7 and Galaxy S7 edge.
-Pre-book the Galaxy S7 and Galaxy S7 edge now and receive a GearVR free!
-Samsung Galaxy S7 is priced at Rs 48,900 (32GB) while Galaxy S7 Edge is available at a price of Rs 56,900 (32GB).
-Keep your cool while gaming. The internal cooling system in Galaxy S7 and Galaxy S7 edge keeps the devices from overheating.
-Galaxy S7 and Galaxy S7 edge enhanced GPU lets you play heavy graphic games and is 60% more powerful than its predecessor.
-Mandeep Bhatia, GM, introduces Samsung Concierge, an exclusive post purchase

Figure: Unstructured data from LexisNexis

The semi-structured data was obtained from multiple sources like fonoApi, GSMArena, phoneArena. FonoApi provided an API for data extraction for each phone model. GSM Arena and Phone Arena were crawled to obtain details about each phone and their feature specifications.

REVIEW		OPINIONS	COMPARE	PICTURES
NETWORK	Technology	GSM / CDMA / HSPA / EVDO / LTE		
LAUNCH	Announced	2015, September		
	Status	Available. Released 2015, September		
BODY	Dimensions	158.2 x 77.9 x 7.3 mm (6.23 x 3.07 x 0.29 in)		
	Weight	192 g (6.77 oz)		
	SIM	Nano-SIM		
DISPLAY	Type	- Apple Pay (Visa, MasterCard, AMEX certified)		
	Size	LED-backlit IPS LCD, capacitive touchscreen, 16M colors		
	Resolution	5.5 inches (~67.7% screen-to-body ratio)		
	Multitouch	1080 x 1920 pixels (~401 ppi pixel density)		
		Yes		

Fig: Phone Arena sample page



DESIGN		COMPARE
Device type:	Smart phone	
OS:	iOS (10.x)	
Dimensions:	 5.44 x 2.64 x 0.28 inches (138.3 x 67.1 x 7.1 mm)	
Weight:	 4.87 oz (138 g) the average is 5.3 oz (150 g)	

Figure: GSMArena sample page

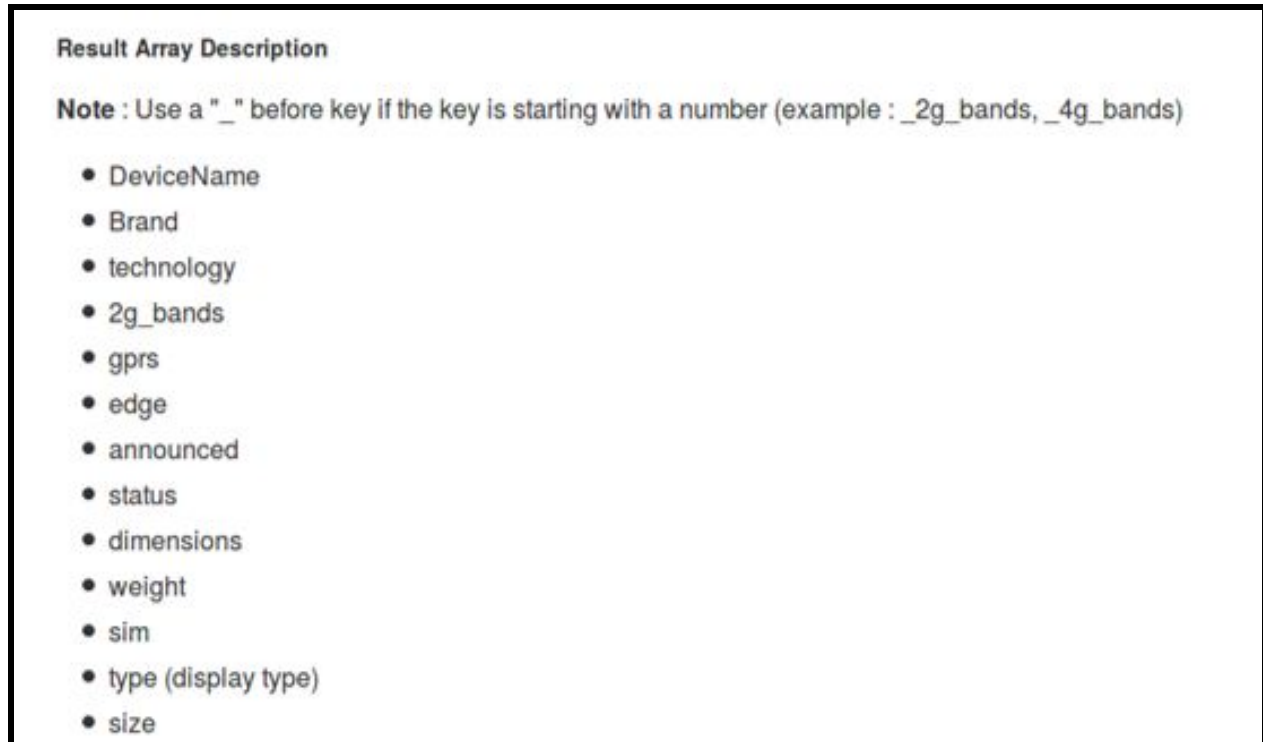


Figure: API provided by fonoApi

3.2 Tools

The comprised tool set we have used is depicted in the figure below. Data layer is comprised of all the data sources. The data was converted to structured form and saved in SQLite database. We used Python 2.7 for all the processing including the data collection, data processing, query processing, etc. The Python package - Natural Language Tool Kit (NLTK) was used for processing of the data and identifying the features and sentiment analysis. Finally, the UI was created using jQuery and Bootstrap packages mainly. Flask was used for creating the server and hosting all the pages.



Figure: Tools used in the different layers of the application

3.3 Feature Extraction and Sentiment Analysis

We used two sets of feature dictionaries - dynamic features and static features for this process. The dictionaries were obtained from the data sources described in the previous section. The documents were split into sentences and phrases by the text extractor. These sentences were then processed to identify the features present in them. The words were then passed through a step where we did stemming and lemmatization and then tagged using a POS tagger present in the NLTK package. Once all the words in the sentences were tagged, the top occurring nouns in those sentences were extracted. This created a list of the top dynamic features present in all the documents.

The next part was creating the static feature dictionary. We used the data from PhoneArena, GSM Arena and fonoAPI which gave us a list of features present in all the phones. The next challenge was to map the dynamic and static features so that we can group together identical features and also help in identifying the feature present in the tweets and the documents. This was achieved by using the similarity function available in the NLTK package which helped in aggregating all the features into a set of static features.

After obtaining all the features, we identified the sentences which were present for all the features and used the sentiment analyser to calculate the sentiment score for those sentences. Hence, for each feature we had a count measure and a sentiment score.

A similar method was used for the tweets where we identified the feature for each tweet and calculated its sentiment score. At the end of the analysis, we had a sentiment score and count measure for each of the features from the twitter data. This process is described in the figure below.

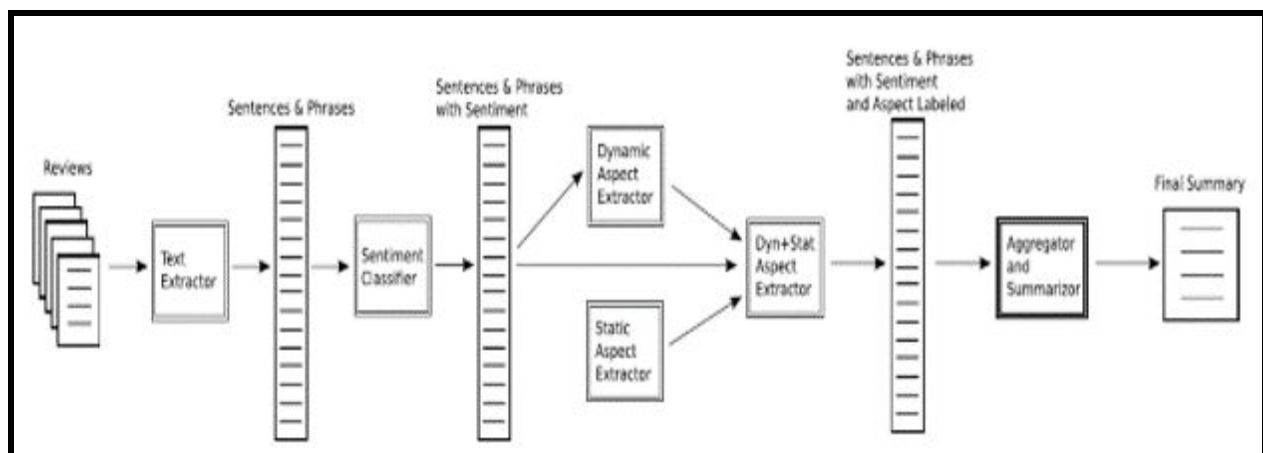


Figure: Aspect based Sentiment Analysis process

3.4 Truth Tables

We tabulated the data obtained through both the data sources. We formed the truth tables for the feature frequency and the corresponding sentiment associated with each feature. We do this for data feed from Twitter and Lexis Nexis.

Below are the snippets of the truth tables.

Note: This is just a representation. The actual truth tables were much larger.

id	speaker	battery_capacity	dimensions
1	Earpiece; Loudspeaker	2915 mAh	6.22 x 3.06 x 0.28 inches (158.1 x 77.8 x 7.1 mm)
2	Earpiece; Loudspeaker	2915 mAh	6.22 x 3.06 x 0.28 inches (158.1 x 77.8 x 7.1 mm)
3	Earpiece; Stereo speakers	1960 mAh	5.44 x 2.64 x 0.28 inches (138.3 x 67.1 x 7.1 mm)
4	Earpiece; Stereo speakers	1960 mAh	5.44 x 2.64 x 0.28 inches (138.3 x 67.1 x 7.1 mm)
5	Earpiece; Loudspeaker	1810 mAh	5.44 x 2.64 x 0.27 inches (138.1 x 67 x 6.9 mm)
6	Earpiece; Loudspeaker	1810 mAh	5.44 x 2.64 x 0.27 inches (138.1 x 67 x 6.9 mm)
7	Earpiece; Loudspeaker	1715 mAh	5.44 x 2.64 x 0.28 inches (138.3 x 67.1 x 7.1 mm)
8	Earpiece; Loudspeaker	1715 mAh	5.44 x 2.64 x 0.28 inches (138.3 x 67.1 x 7.1 mm)

Table: Feature details

phones	battery	camera	memory	performance
iphone_6p	4	11	2	1
iphone7	20	140	18	459
iphone_6	3	50	6	14
iphone_6s	19	52	8	25
galaxy_s7	0	12	0	3
pixel	2	1993	38	52
lg_g5	1	1	0	1

Table: Feature Frequency (Twitter)

phones	color	photo	proximity	compass
iphone_6plus	0	0	0	-0.034483
iphone7	0.002425	0.018724	0	0.001334
iphone_6	0	0.010989	0	0
iphone_6s	0.016873	0.016803	0.037037	0
galaxy_s7	0	0	0	0
pixel	0.011765	0.005839	0	0.002451

Table: Feature Sentiment (Twitter)

phones	color	headphone	design	size	network	pay
iphone_6plus	186	486	1034	2254	478	1416
iphone7	948	6852	3184	1870	438	288
iphone_6	178	262	1360	2980	478	990
iphone_6s	458	1236	1866	2502	604	938
galaxy_s7	98	578	1256	1464	6180	578

Table: Feature Frequency (Lexis Nexis Data)

phones	color	headphone	design	size	network	pay
iphone_6plus	0.060038	0.000377	0.020833	0.011732	0.032153	0.02557
iphone7	0.050496	-0.002332	0.020713	-0.004181	0.013001	0.017165
iphone_6	0.037073	0.008157	0.021759	0.009365	0.031506	0.017868
iphone_6s	0.051069	0.00528	0.029653	0.007394	0.029702	0.028198
galaxy_s7	0.033097	0.017096	-0.048243	0.015873	0.018461	0.02956
pixel	0.034696	0.020564	-0.033643	0.011111	0.024006	0.022202
lg_g5	0.048133	0.015139	-0.029787	0.014273	0.007498	0.029358

Table: Feature Frequency (Lexis Nexis Data)

3.5 Query Processing

We wanted a customizable recommender which considers the budget/ price range and a ranking for a set of fixed features provided. The features were selected based on their frequency obtained from the analysis of the LexisNexis datasets and the twitter content. The user can select the price of the new device which are in the ranges \$600-\$700, \$700-\$800 and \$800-\$900. This decision was made based on the most recent prices of the iPhones. This is the only source of structured data we get from the user for the recommendation system.

After obtaining all these inputs from the user, the query processor identifies the

specifications for each of the features based on their relative ordering. We narrowed down the phones based on the price range obtained from the user for the feature search. For the top half of the features in the prioritized feature list, we identified which phone has the best sentiment and returned its specification. For the remaining features, we selected the specifications randomly from the phones in the same price range. The reason behind such a decision is that giving a top specification for all the features with a price constraint is not a solvable problem and hence, this decision was made. It is not solvable since the individual feature price is not available publicly.

4. OBSERVATIONS

After performing the sentiment analysis, we have all the features and the associated sentiment scores for each. We recognize the most popular features as these are the most influential ones in decision making.

Q. How did we obtain the most popular features?

- The popularity of a feature is quantified by how frequently that feature was mentioned in the data.
- We formed a frequency table of the features and their mention in all of the data sources combined i.e tweets, reviews, news articles etc.
- From this feature frequency table, we picked up the top 10 features and these features form our dataset for the most sought after features in a phone.

Given below is the list and the corresponding popularity of each feature expressed as a percentage of the total count of these 10 features: (In descending order)

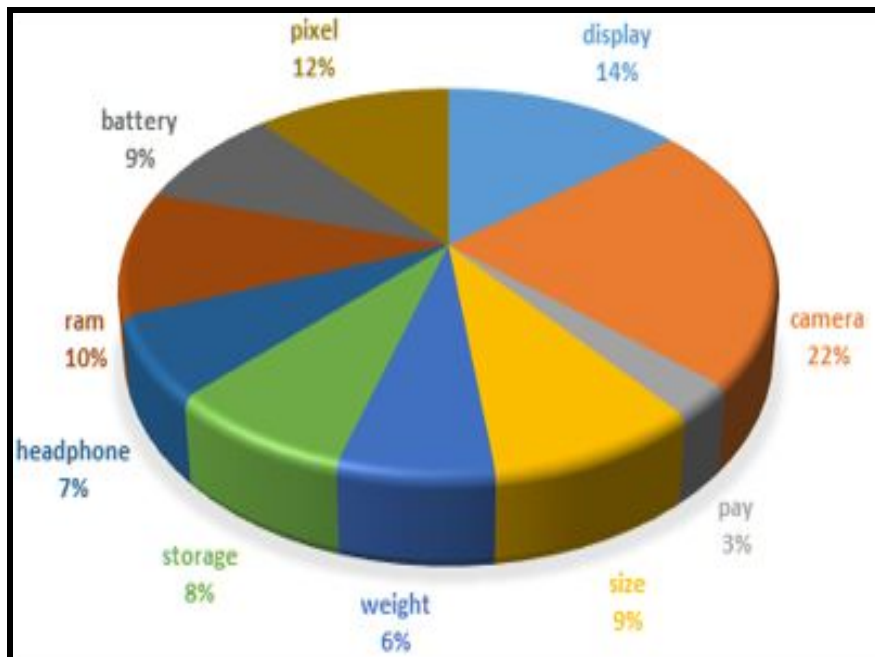


Figure: Feature popularity distribution

Feature	%
camera	22
display	14
pixel	12
ram	10
battery	9
size	9
storage	8
headphon e	7
weight	6
pay	3

Table: Feature popularity distribution

Here are some interesting observations based on the analysis of the feature sentiment:

Camera: Presented below is a graph of the sentiment score associated with each phone for the camera feature. According to our observations, Galaxy S7 has the best camera followed by LG G5 and so on. Surprisingly, the camera for Google Pixel turned out to be the least popular one.

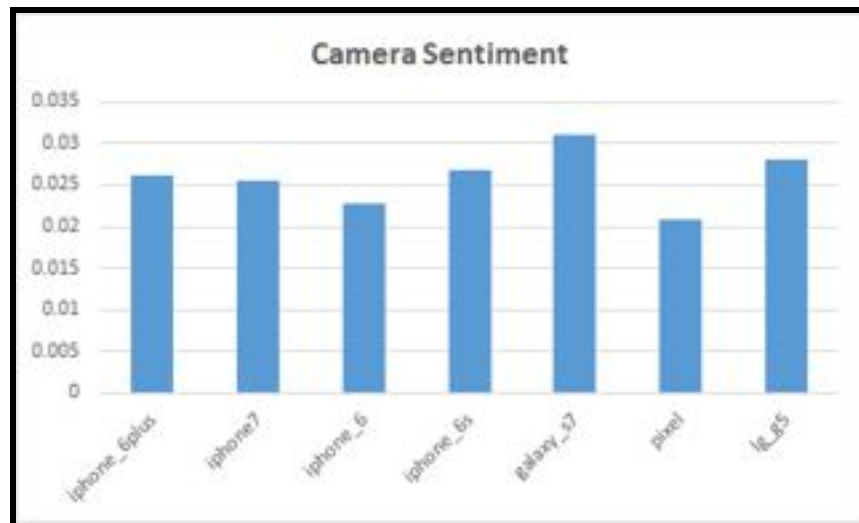


Figure: Sentiment scores for the feature 'camera'

Headphone: The graph below shows the sentiment score associated with each phone for the headphone feature. According to our observations, Google Pixel has the best camera followed by Galaxy S7 and so on. Based on the recent uproar with the missing headphone jack for Iphone 7, it is no surprise that Iphone 7 had a negative sentiment.

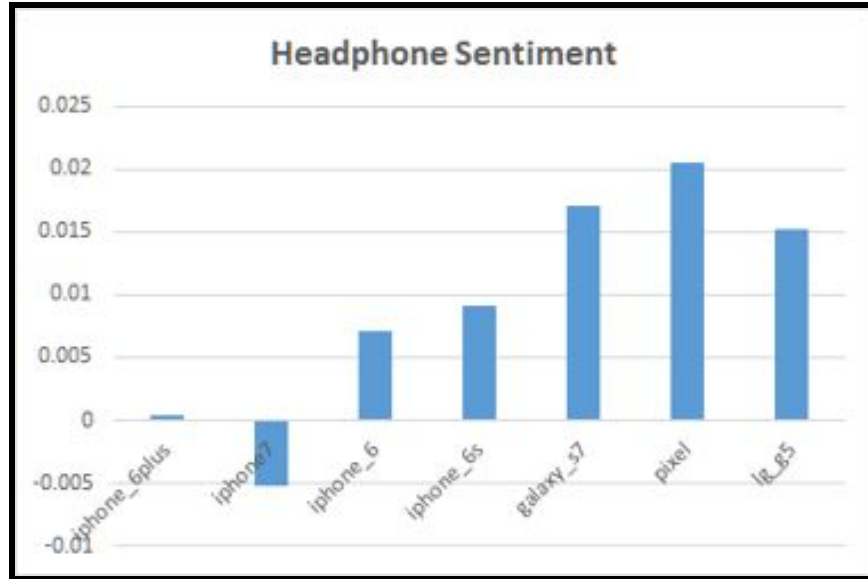


Figure: Sentiment scores for the feature 'headphone'

5. CONCLUSION

We were able to successfully develop an application for the given problem statement. During the course of the project we learnt how to break down a problem to perform meaningful analysis, how to work on the real world data that is naturally very unstructured. We also got a chance to work with various tools and build a recommender system. We learnt the essentials of critical thinking and gave the world a cool application.

6. FUTURE WORK

- Using an exhaustive dataset in terms of more phones, tweets and reviews.
- Ability to identify the cost associated with each feature that will help us with making better recommendations based on budget range.
- A better classifier for identifying/tagging the sentences with the feature.
- A good evaluation system to gauge our results.

7. REFERENCES

1. Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." *Foundations and trends in information retrieval* 2.1-2 (2008): 1-135.
2. Amarouche, Kamal, Houda Benbrahim, and Ismail Kassou. "Product Opinion Mining for Competitive Intelligence." *Procedia Computer Science* 73 (2015): 358-365.
3. "Sentiment Analysis". *En.wikipedia.org*. N.p., 2016. Web. 1 Dec. 2016.
4. Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." *Proceedings of the workshop on languages in social media*. Association for Computational Linguistics, 2011.
5. Pang, Bo, and Lillian Lee. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2004.

8. APPENDIX

GitHub link: https://github.com/moharnab123saikia/sales_opportunity