University of CINCINNATI

TEXT MINING ON USER PHONE REVIEWS TO STUDY USER OPINIONS

BANA 7047 – Data Mining II

Section 001 - Group 3

By

Abhijith Antony (M13433446)

Hridhay Mehta (M13474723)

Sanjay Jayakumar (M13419548)

Vishnu Vijayakumar (M13255870)

Abstract

Amazon is one of the most popular online marketplaces used by both individual as well as businesses, and the site is available in many different countries and languages. Not only is Amazon the most valuable internet retailer in the world, it is the most valuable retailer period, having surpassed WalMart back in 2015. (Source: https://ecommerce-platforms.com/glossary/amazon) thus underlining its importance in the world marketplace.

Having a customer base of over 150 million in the US alone, the large volume of data that is generated from the Amazon app & website in the form of customer reviews and buying patterns / purchase decisions etc. are of enormous value and are treated ass valuable insights to make data driven decisions.

In this project, we deploy topic modeling and sentiment analysis to identify the major concerns that mobile phone customers share in the form of reviews and ratings. We proceed to do this through different unsupervised learning approaches like Latent Dirichlet Allocation & Non-Matrix Factorization. The dataset we use for the study consists of ~68k reviews for over 720 phones / brands sold on Amazon. The results obtained look promising and we can distinctly point out the negative and positive aspects of the brand / product the customer identifies.

An added interest is to use predictive modeling techniques like logistic regression, Support Vector Machine, Random Forest and Naïve Bayes classifier to predict user ratings from their text reviews. Being a topic of academic interest for us, the models have F1 scores in the range 0.62 – 0.69.

Keywords: Amazon, text mining, topic modeling, sentiment analysis, predictive modeling

An online version of our project is hosted on Google Colab.
Please follow the link https://colab.research.google.com/drive/1WolAb0Al-9LwdQp10THen-38VLlzpjfb#scrollTo=YvKBs7h5rjwL&uniqifier=1 and request access.

Contents

Chapter 1 Introduction

1.1 Background

Amazon receives huge amount of reviews on various Brands and products it sells, and it needs sophisticated machine learning techniques to understand customer feedback which is usually in the form of unstructured text data. For understanding customer concerns and to be able to do it accurately with least manual intervention, its necessary to employ machines to understand data and make sense of it. We want to employ techniques like Topic Modeling, Sentiment analysis and various machine learning algorithms for user rating predictions to make sense of these large unstructured datasets.

1.2 Goal of the Thesis

In this project, we will be exploring over 68k customer reviews of over 720 mobile phones posted on Amazon.

We aim to take a two-pronged approach in this project. One, to use the techniques of topic modeling to point out the top positive and negative aspects of purchase that the users associate with a brand / product based on their reviews. In this respect, we will focus on the application of LDA (Latent Dirichlet Allocation) and NMF (Non-Matrix Factorization) towards achieving this goal and then comparing the outputs obtained by these two techniques

The second prong of the project will be pointed at creating a predictive model to predict user ratings by exploring logistic regression, Support Vector Machine, Random Forest and Naïve Bayes and arrive at the best model to do the same.

Chapter 2 Exploratory Data Analysis

2.1 Data Gathering

In this project we are working with 2 files:

- reviews.csv: This file has around 68K reviews with 8 feature variables for all brands of phones. There are 2 numerical variables rating and helpfulVotes.
- items.csv: Pre-scraped data for 720 phone items from amazon with 10 feature variables. There are 4 numerical variables: rating, totalReviews, price and original price.

We merged both these files with the common column 'asin'.

2.2 Data Cleaning

It is important that we have the data in clean and standard format before proceeding to further analysis. Data cleaning steps followed in the project:

- First step is to do the basic cleaning processes like converting text into lowercase, removing punctuations. This was done using regular expressions (re) in python.
- Next, we checked for null values in the dataset. For reviews.csv we saw a significant amount of NA's for column helpfulVotes which was then removed.
- We also removed stop words; these are usually the most used words in English. Removing such words will help us focus on other important words in the text.
- There is a need to tokenize data, which essentially means to split the text into smaller pieces.
- We also performed stemming and lemmatization, which are text normalization techniques. We used the NLTK package in Python to perform the same. Stemming is the process of reducing the words to their root forms, like mapping a group of words to the same step. Although lemmatization ensures that the root word is a valid word in English language unlike stemming.

2.3 Visualization

Visualization is a very strong medium to understand the data. We generated multiple plots to understand the nature of overall and brand wise user ratings before venturing into Sentiment analysis and topic modeling. *Figure 1* gives a brand wise outlook on the user ratings. The highest deviation in ratings due to missing reviews is for one-plus phones at 0.63. Samsung, Nokia, Motorola and Apple have less than 0.2 rating deviation.

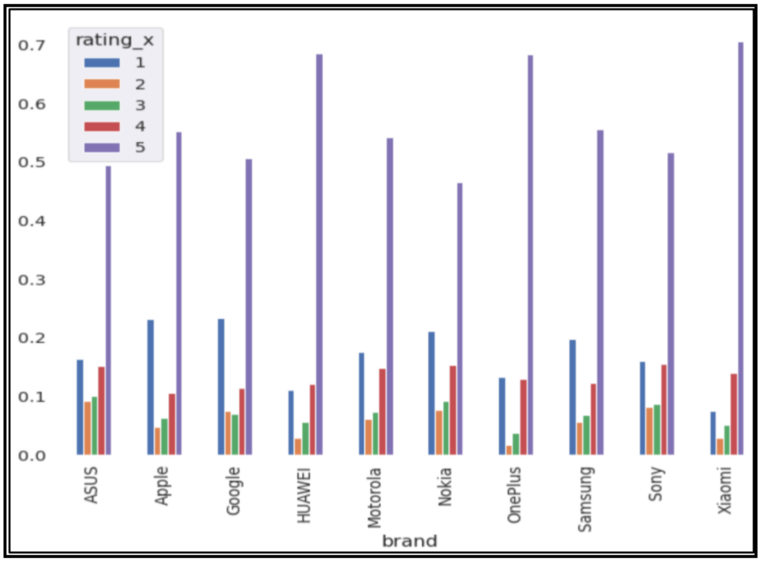


Figure 1: Average user ratings per Brand

We can see from *figure 2* that out of all the 68K reviews, more than 70% of them have been rated as either 4 or 5. Hence we see a bi-modal distribution of ratings for all the brands in the figure.

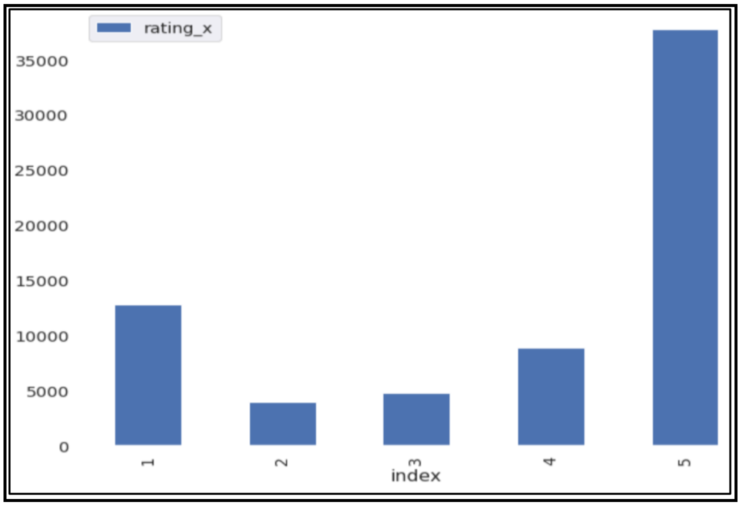


Figure 2: Overall user ratings

Chapter 3 Sentiment Analysis

3.1 Sentiment Analysis Methods

Sentiment Analysis is contextual mining of text which identifies and extracts subjective information in source material and helps a business understand the social sentiment of a brand. Here we are dealing with user reviews of multiple brands and hence this technique is very essential to understand the customer sentiment.

Input to a Sentiment analysis technique is a corpus, which is basically a collection of words where order matters.

Methods used for Sentiment analysis:

- **TextBlob**: TextBlob is a part of NLTK library in python. The output of sentiment analysis is a sentiment score ranging from -1 to 1 (which is polarity) indicating how positive or negative they are and a subjectivity score of 0 to 1 where 0 indicates a fact and 1 indicates an opinion. TextBlob finds all the words and phrases that it can assign a polarity and sensitivity to and averages them all together. Finally, each phone is assigned one polarity and one subjectivity scores.

- **Vader**: Vader (Valence Aware Dictionary and Sentiment Reasoner) is a part of vaderSentiment library in Python for applying sentiment analysis techniques. The output of Vader method is a compound score metric which is calculated by summing the valence scores of each word in the lexicon, adjusted according to the rules, and normalized to be between -1 (most extreme negative) and 1 (most extreme positive).

3.2 Interpreting Results

For sentiment analysis we have tried to use standardized thresholds to classify sentences as either positive or negative. This is required since we saw that the data has imbalanced classes of rating with 4,5 taking up about 67% of the data. Hence, we need a higher threshold to classify 4,5 ratings as positive. After trying multiple threshold values, we see that for 4,5 0.5 cut works the best and for the rest 0 cut off. Below are the results obtained when keeping the threshold at 0 and 0.5 and both Vader and TextBlob sentiment analysis techniques.

Text	Cutoff	Sentiment	Rating				
			1	2	3	4	5
body	0.5 for 4,5 0 else	**0**	7285	1866	1517	1001	11380
		1	5458	2049	3235	7823	26372
	0.5	**0**	11221	3020	3183	3739	11380
		1	1522	895	1569	5085	26372
	0	**0**	9500	2369	2083	1768	5557
		1	3243	1546	2669	7056	32195
title	0	**0**	11340	3150	3220	3771	14643
		1	1403	765	1532	5053	23109
	0.5	**0**	12388	3691	4283	6287	21439
		1	355	224	469	2537	16313

Table 1: Rating Summary

Brand		cutoff = 0						cutoff = 0.5					
		TextBlob			VADER			TextBlob			VADER		
		sentiment	+	-	sentiment	+	-	sentiment	+	-	sentiment	+	-
ASUS	title	0.306	132	119	0.207	128	123	0.306	92	159	0.207	76	175
Apple		0.245	2462	2683	0.187	2441	2704	0.245	1513	3632	0.187	1363	3782
Google		0.269	1989	1798	0.199	1895	1892	0.269	1219	2568	0.199	1233	2554
HUAWEI		0.348	1243	982	0.281	1262	963	0.348	855	1370	0.281	833	1392
Motorola		0.265	4276	4604	0.216	4271	4609	0.265	2697	6183	0.216	2720	6160
Nokia		0.217	2579	3336	0.175	2583	3332	0.217	1524	4391	0.175	1622	4293
OnePlus		0.383	210	137	0.288	207	140	0.383	145	202	0.288	153	194
Samsung		0.237	15045	18584	0.192	15022	18607	0.237	9200	24429	0.192	9328	24301
Sony		0.249	1572	1624	0.212	1571	1625	0.249	898	2298	0.212	1010	2186
Xiaomi		0.346	2350	2061	0.271	2379	2032	0.346	1653	2758	0.271	1500	2911
ASUS	body	0.253	198	53	0.800	186	65	0.253	40	211	0.450	148	103
Apple		0.238	3377	1768	0.774	3265	1880	0.238	1053	4092	0.304	2276	2869
Google		0.244	2904	883	0.800	2638	1149	0.244	657	3130	0.371	2095	1692
HUAWEI		0.301	1624	601	0.806	1594	631	0.301	561	1664	0.457	1293	932
Motorola		0.271	6742	2138	0.781	6387	2493	0.271	1885	6995	0.405	4870	4010
Nokia		0.238	4333	1582	0.802	4052	1863	0.238	1086	4829	0.373	3076	2839
OnePlus		0.294	281	66	0.808	264	83	0.294	77	270	0.488	216	131
Samsung		0.277	24076	9553	0.779	22912	10717	0.277	8073	25556	0.362	17227	16402
Sony		0.269	2526	670	0.809	2357	839	0.269	613	2583	0.448	1901	1295
Xiaomi		0.280	2921	1490	0.806	2899	1512	0.280	1005	3406	0.397	2204	2207

Table 2: Brand Summary

We also compared the results for sentiment analysis based on two different columns, one its title and the other it's body text (*Table 2* and *Table 3*).

Chapter 4 Topic Modeling

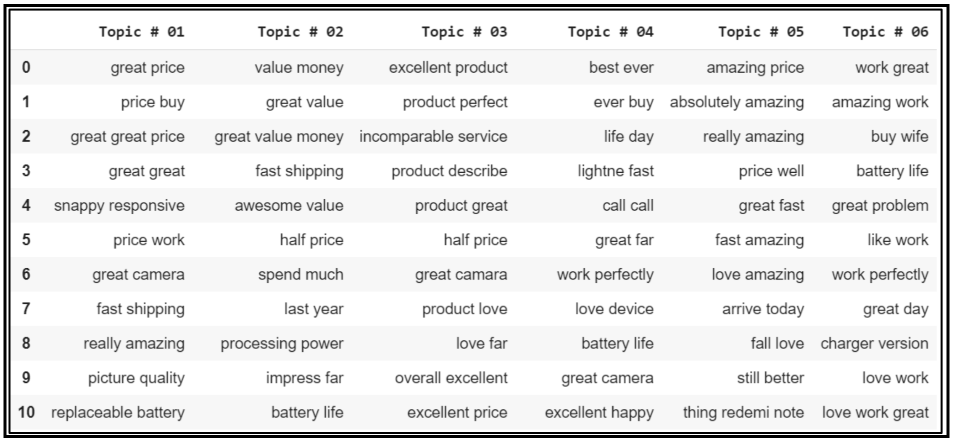
4.1 Topic Modeling Methods

Topic modeling is a technique used to extract meaningful information from vast amounts of data. In this project we are trying to label different topics among all the cell phone reviews in order to provide business recommendations to sellers.

Input to the topic modelling technique is a document-term Matrix: It is a matrix where the rows are different documents and columns are different terms and values in the matrix are the word counts. Each topic will consist of a bag of words not necessarily ordered.

We are going to implement topic modeling by using a python library called Gensim. This package utilizes a topic modeling technique called Latent Dirichlet Allocation (LDA). This technique aims to find the hidden probability distributions where every document is a probability distribution of topics and every topic is a distribution of words. We give the document-term matrix, number of topics and number of iterations as input to the Gensim LDA process. Gensim will go through the process of finding the best word distribution for each topic and best topic distribution for each document.

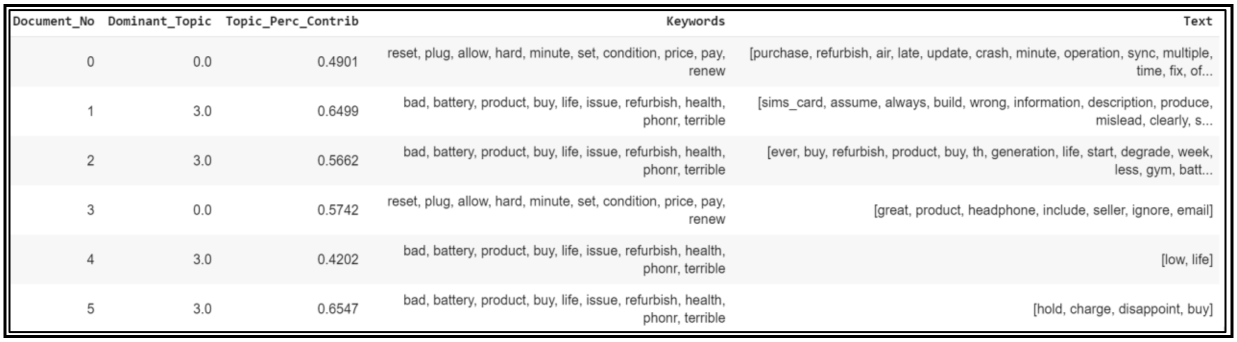
We have also performed Nonnegative-Matrix Factorization, input to this algorithm is Document-Term matrix and number of topics. The output we receive from the algorithm are two non-negative matrices of the original words by K-topics and those k topics by the m original documents. Below are the results we received for Xiaomi for positive feedbacks using Nonnegative-Matrix Factorization.



	Topic # 01	Topic # 02	Topic # 03	Topic # 04	Topic # 05	Topic # 06
0	great price	value money	excellent product	best ever	amazing price	work great
1	price buy	great value	product perfect	ever buy	absolutely amazing	amazing work
2	great great price	great value money	incomparable service	life day	really amazing	buy wife
3	great great	fast shipping	product describe	lightne fast	price well	battery life
4	snappy responsive	awesome value	product great	call call	great fast	great problem
5	price work	half price	half price	great far	fast amazing	like work
6	great camera	spend much	great camara	work perfectly	love amazing	work perfectly
7	fast shipping	last year	product love	love device	arrive today	great day
8	really amazing	processing power	love far	battery life	fall love	charger version
9	picture quality	impress far	overall excellent	great camera	still better	love work
10	replaceable battery	battery life	excellent price	excellent happy	thing redemi note	love work great

Figure 3: Positive Topic results from non-negative matrix factorization for Xiaomi

Output of topic modeling will be top words in each topic highlighting themes across various user reviews. From the output we can interpret results and see if the bag of words in each topic makes sense.



Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0.0	0.4901	reset, plug, allow, hard, minute, set, condition, price, pay, renew	[purchase, refurbish, air, late, update, crash, minute, operation, sync, multiple, time, fix, of...
1	3.0	0.6499	bad, battery, product, buy, life, issue, refurbish, health, phonr, terrible	[sims_card, assume, always, build, wrong, information, description, produce, mislead, clearly, s...
2	3.0	0.5662	bad, battery, product, buy, life, issue, refurbish, health, phonr, terrible	[ever, buy, refurbish, product, buy, th, generation, life, start, degrade, week, less, gym, batt...
3	0.0	0.5742	reset, plug, allow, hard, minute, set, condition, price, pay, renew	[great, product, headphone, include, seller, ignore, email]
4	3.0	0.4202	bad, battery, product, buy, life, issue, refurbish, health, phonr, terrible	[low, life]
5	3.0	0.6547	bad, battery, product, buy, life, issue, refurbish, health, phonr, terrible	[hold, charge, disappoint, buy]

Figure 4: Formatted LDA output from Gensim package

4.2 Interpreting Results

4.2.1 Coherence Scores

Topic Coherence scores measures score of a single topic by measuring the degree of semantic similarity between high scoring words in the topic. *Figure 4* provides the optimum number of topics to be extracted using LDA. We pick the number of topics at the highest coherence scores.

The below plot is the coherence score chart for positive review sentiment for Xiaomi. Although we see that the peak of the chart is after 30, this is mostly due to the repeating words in the new topics. Hence, we would consider the first peak at ~8 to be out optimum number of Topics to be modelled.

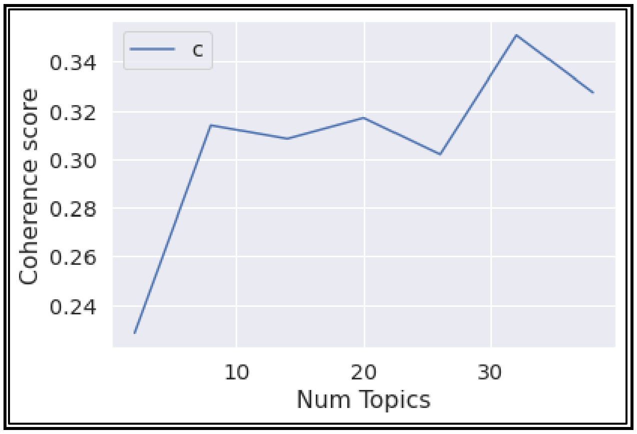


Figure 5: Coherence Scores (Xiaomi positive)

4.2.2 Inferences

Below are the results of Topic modeling for Brands Samsung, Apple and Xiaomi for both positive and negative sentiment. Based on the topics generated from the LDA technique we will analyze the bag of words in both positive and negative topics to identify the good or bad qualities about the product.

Interpretation of results obtained for Xiaomi:

Positive:

- From the iterations carried out for different number of topics, it was observed that the when the number of topics was set to 8, the algorithm was able to uncover some interesting aspects that were not obvious otherwise.
- The consumers were in general, appreciative of the value for money these phones offered to the customer which is one of the main reasons Xiaomi has seen a steady growth over the years. Customers were happy about the camera provided in most the Xiaomi products. People were very pleased with the picture reproduction quality offered by these phones at the competitive prices at which they were sold in the market. The touch system on these smartphones were also appreciated highly by the customers as being very responsive. One interesting aspect that was uncovered was the popularity of the Redmi note pro series among the customers. Most the happy customers were supposedly proud owners of this device from Xiaomi.

Negative:

- From the iterations carried out for different number of topics, it was observed that the when the number of topics was set to 8, the algorithm was able to uncover some interesting aspects that were not obvious otherwise.
- For Xiaomi, the consumers were in general unhappy about the call quality. There might have been several cases of call drops. Customers also raised issues concerning the fingerprint reader on the Xiaomi phones along with some general start up issues some models were facing. Customers were also found to be raising concerns about the security apps installed on the Xiaomi phones. One interesting topic that was generated was concerning the sellers. The phones bought through these sellers might have had several issues. As Xiaomi does not rely on retail stores unlike apple or Samsung the company might find it beneficial to take measures when choosing the authorized sellers to market their phones to customers in different markets.

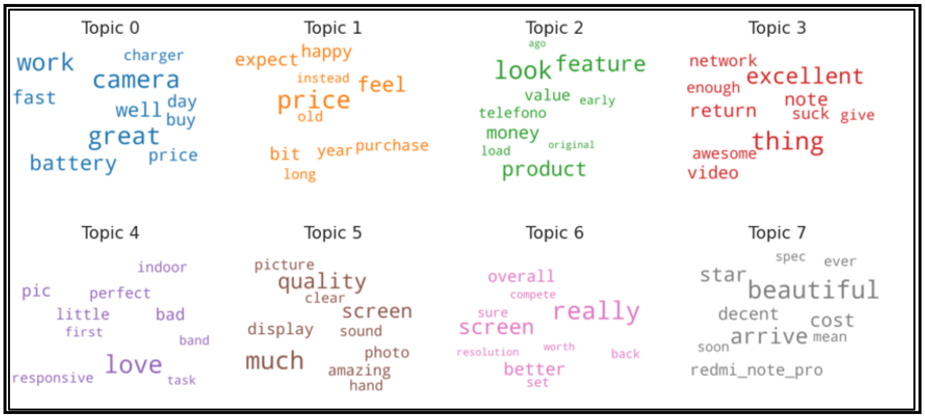


Figure 6: Xiaomi Top 8 positive topics, Word Cloud

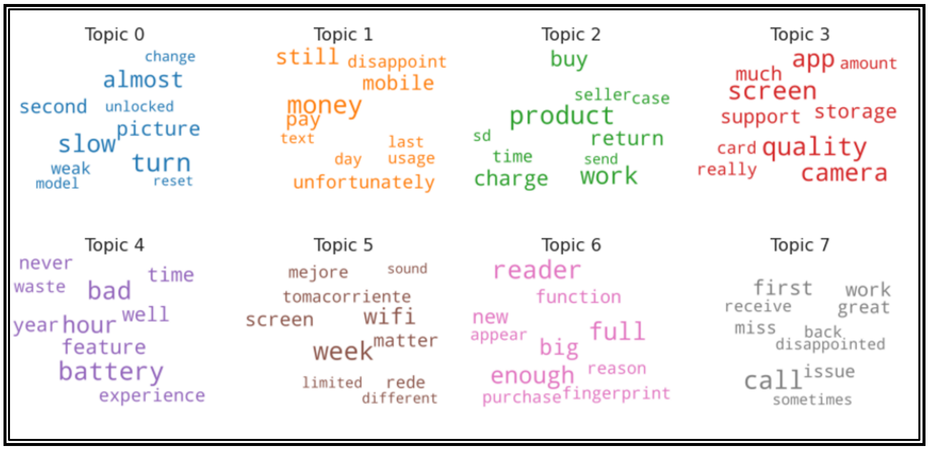


Figure 7: Xiaomi Top 8 negative topics, Word Cloud

Interpretation of results obtained for Samsung:

Positive:

- The users heavily talk about the screen and are happy with the looks. Samsung Galaxy flagship phones are super according to users and seen to have fast performance. Two topics are seen for budget phones talk where users look for a little more quality in speakers and battery but are still content with the overall price. The major liking of Galaxy Note series seems to be the screen. Most of its phones seem to have a good photo quality.

Negative:

- Starting from 2 topics we went on increasing topics to discover the various issues faced by the brand. 10 topics seemed to divide the issues in a distinguishable manner. The most dominant topic was concerning receiving calls followed by issues with screen likely due to drops. Product warranty and boxing issues were also prevalent. The Galaxy Note phone series was also a topic which talked about cases and storage being a problem. Battery charging seemed to be a high reason for returns and some had multimedia issues such as video quality being poor and bad speakers.

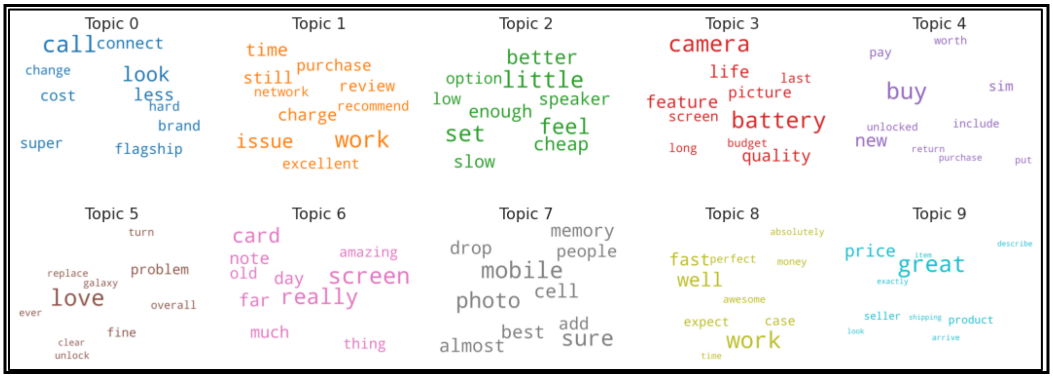


Figure 8: Samsung Top 10 positive topics, Word Cloud

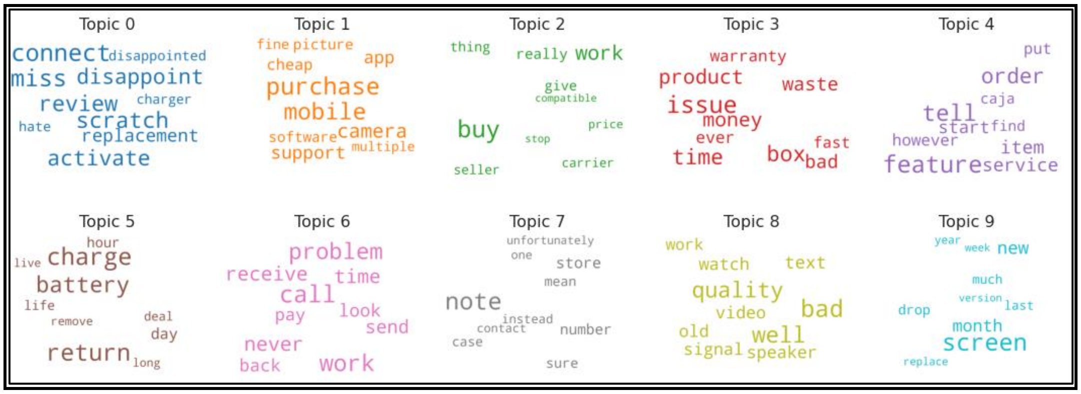


Figure 9: Apple Top 10 negative topics, Word Cloud

Interpretation of results obtained for Apple:

Positive:

- Many Apple phones are sold refurbished on Amazon, and their condition seems to be pretty good which shows up as happy users. The users seem to be fond of the cost, condition and the shipping and it looks like Amazon is doing a good job to sell these phones. Screen is seen to appear often with the word scratch which makes sense as well and is a major topic in high rated reviews. The multimedia keywords appear with excellent keyword, so users seem to be happy with the overall product.

Negative:

- From the iterations carried out for different number of topics, it was observed that the when the number of topics was set to 8, the algorithm was able to uncover some interesting aspects that were not obvious otherwise.
- For Apple, the consumers were in general unhappy about the durability of the phones. The screen (damaged during the shipping process/other internal issues) speaker, battery and charging cable were some of the product aspects the customers were in general disappointed about, which were captured in most of the topics. Being a premium phone company, the price of the phones was also listed as a major issue. Apart from these, another one interesting aspect e customers complained about was related to over-heating

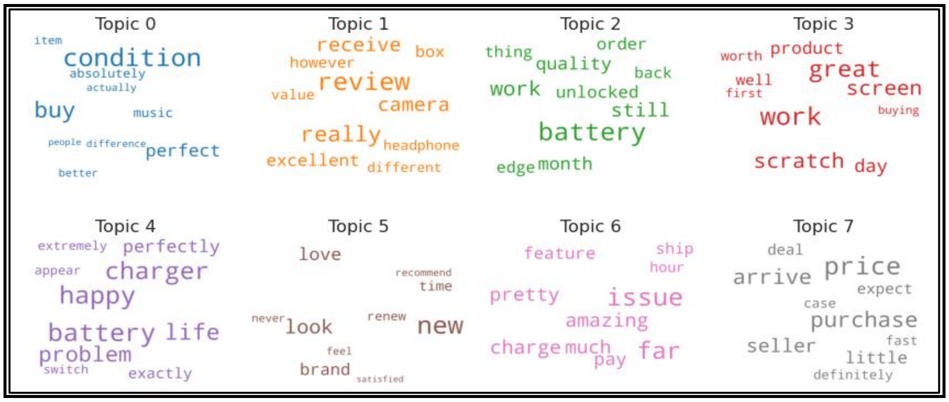


Figure 10: Apple Top 8 positive topics, Word Cloud

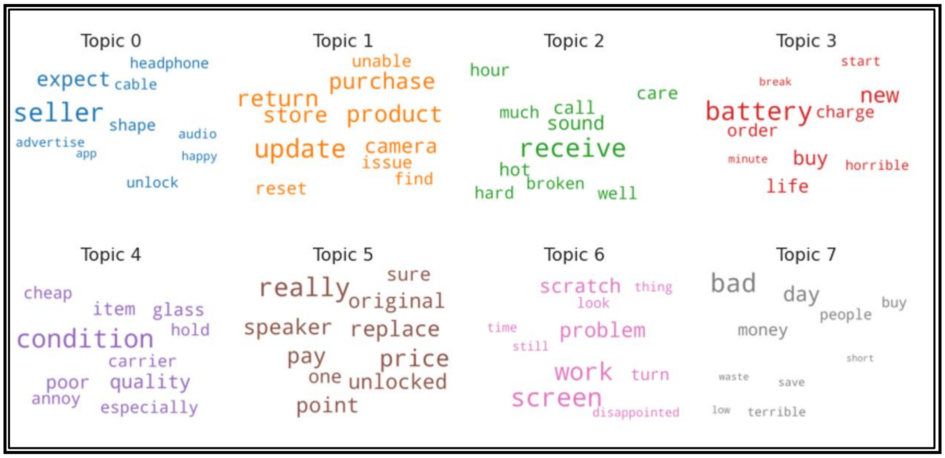


Figure 11: Apple Top 8 negative topics, Word Cloud

Chapter 5 User Rating Prediction

5.1 User Rating Prediction Methods

The rise in E — commerce, has brought a significant rise in the importance of customer reviews. There are hundreds of review sites online and massive amounts of reviews for every product. Customers have changed their way of shopping and according to a recent survey, 70 percent of customers say that they use rating filters to filter out low rated items in their searches.

The ability to successfully decide whether a review will be helpful to other customers and thus give the product more exposure is vital to companies that support these reviews, companies like Google, Amazon and Yelp!

Here we will try to predict ratings based on customer's text reviews using different models and compared their performance by their accuracies. We have built Logistic regression, Naïve Bayes, SVM, Random Forest models. We would be following below steps to predict ratings based on these models:

- Balance the data: One of the complications we see is that the dataset is unbalanced with 67% of the overall ratings as 4,5. We will get biased results if we train the system on this unbalanced data. So, the first step before training a model is to balance a dataset by removing the overrepresented samples of ratings. Although, for the current analysis we have not done the balancing, however we will consider the bias while interpreting the results.
- Convert text data into TF-IDF vectors: Term Frequency Inverse document frequency vectors normalizes the count of each words in each text by the number of times the word occurs in all the texts. This is based on the theory that usually high frequency words are less important.
- Split the data into a training and test set: We will split the dataset into test and training data in order to check the out of sample accuracy of our trained models.
- Classify the text data using different models: Classifiers Used: Logistic regression, Naïve Bayes, SVM, Random Forest
- Evaluate the performance of each of the models using precision, recall and accuracy.

5.2 Model Performance Evaluation

Model Performance evaluations of all the classifiers will be based on below Metrics:

- Precision: It is also called the positive predictive value and is the fraction of relevant instances among retrieved instances. A model that produces no false positive has a precision of 1.
- Recall: It is also called sensitivity and is the fraction of the total amount of relevant instances that were retrieved. A model that produces no false negatives has a recall of 1.
- F1 score: In certain cases we want to maximize either recall or precision at the expense of other metric, however in most cases we want to find an optimal blend of precision and recall, and F1 score is used to combine the same. This is the harmonic mean of precision and recall.

From *table 3* in summary interpretation of results for various classifiers:

	Logistic Regression				**SVM**			
Ratings	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**
1	0.68	0.71	0.69	3844	0.7	0.56	0.62	3844
2	0.24	0.24	0.24	1122	1	0	0	1122
3	0.27	0.2	0.23	1409	0.42	0	0.01	1409
4	0.34	0.2	0.25	2606	0.22	0	0	2606
5	0.8	0.9	0.85	11415	0.66	0.99	0.79	11415
Accuracy			**0.69**	**20396**			**0.66**	**20396**
	Naïve Bayes				**Random Forest**			
Ratings	**Precision**	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**
1	0.77	0.49	0.6	3844	0.61	0.33	0.43	3844
2	0	0	0	1122	0.88	0.07	0.13	1122
3	0	0	0	1409	0.92	0.07	0.13	1409
4	0.33	0	0	2606	0.77	0.06	0.11	2606
5	0.63	1	0.77	11415	0.62	0.97	0.76	11415
Accuracy			**0.65**	**20396**			**0.62**	**20396**

Table 3: Model Comparison

Logistic Regression

The accuracy for the logistic regression model was found to be 68.68%. The precision for class 1 is 0.68 which means out of all class 1 predictions the number of correct class 1 predictions is 0.61. The recall is found to be 0.71 which means the proportion of correctly predicted class 1 out of total number of actual class 1 cases is 0.71. The F1 score is a harmonic mean of these two figures and the micro avg is the simple mean of f1 scores or precision recall figures over all the classes. Here when we calculate the micro average scores, we gave same weightages to all the classes. In weighted scores, we weight the scores from each class by the number of samples from that class. The micro F1 score is same as the overall accuracy of the model. The micro precision score and recall scores are the same. Since our data is highly imbalanced, it is better to look at the weighted f1 score to compare the models. It can be observed that the logistic regression was able to classify the class 1 and class 5 models accurately, but the other class predictions were abysmal. Thus, we need to explore other models which can overcome this potential issue.

SVM

The accuracy of the model was obtained as 66.19% which is same as the micro F1 score for the model. The precision for the different classes in SVM is much better than the logistic regression model. However, it comes at the expense of the overall model accuracy. But considering the dataset at hand, and the fact that the reviews would be in general skewed to have a positive sentiment, SVM would be better choice as when compared to Logistic regression.

Naïve Bayes

Another classifier we explored was the Naïve Bayes classifier. The same idea and pipeline module were implemented in calculating the model performance. Even though at first glance, the model seems okay, it must be noted that the model performed the worst in predicting the class 2,3 and 4 ratings.

Random Forest

The final model tested in the analysis was the Random Forest model. A slight change was made in how the model was built. While constructing the document term matrix using the TF-IDF model, both bigrams and unigrams were considered while building the Document Term Matrix. Rather than using a count vectorizer, a TF-IDF vectorizer was directly utilized in the model to understand the effects on the model accuracy. It can be observed that the random forest model with the ngram technique was able to improve the precision scores for all the classes. For example, for class 3 ratings, out of all the predicted class 3 ratings, 92% of them were correct. However, this reduced the recall score for all the classes consequently. For class 3 ratings, out of the total actual number of class 3 cases, only 7% were predicted correctly by the model. The model has a micro average F1 score of 0.62 which corresponds to the overall accuracy of the model.

Chapter 6 Conclusion

6.1 Summary

We have approached this project from two angles.

One, a business / real world perspective where we use techniques of topic modeling to point out the top positive and negative aspects of brands based on user reviews. We focused on the application of LDA (Latent Dirichlet Allocation) and NMF (Non-Matrix Factorization) towards achieving this goal.

The second aspect of the project was academic in nature. We aimed to predict the user ratings based on the user reviews by deploying logistic regression, Support Vector Machine, Random Forest and Naïve Bayes.

70% of the reviews were skewed to 4 and 5 rating. A bi-modal distribution of ratings was observed for all the brands. This leads us to sentiment analysis.

We use standardized thresholds to classify sentences as either positive or negative. This is imperative given the skew of the ratings. So, we take a higher threshold of 0.5 to classify 4,5 ratings for the rest 0 cut off. We use both TextBlob and Vader for this purpose.

Next, we proceed to do topic modeling, which is a technique used to extract meaningful information from vast amounts of data. We try to label different topics among all the cell phone reviews in order to provide business recommendations to sellers. LDA and NMF are the two techniques we have used to do topic modeling.

Input to the topic modelling technique is a document-term matrix, where the rows are different documents and columns are different terms and values in the matrix are the word counts. Each topic will consist of a bag of words not necessarily ordered. The output of topic modeling will be top words in each topic highlighting themes across various user reviews. From the output we can interpret results and see if the bag of words in each topic makes sense.

We perform topic modeling for 3 major players in the Smartphone segment: Apple, Samsung and Xiaomi. The detailed findings are explained in the previous sections.

Once we complete topic modeling, we proceed to expand our academic interests by creating a predictive model that gives the user ratings based on their reviews. The ability to successfully decide whether a review will be helpful to other customers and thus give the product more exposure is vital to companies that support these reviews, companies like Google, Amazon and Yelp! We use Logistic Regression, Naïve Bayes, SVM and Random Forest to create such a model.

6.2 Future work

Some of the steps that we take to further our study:

- Cleaning is a very important aspect especially for sentiment analysis and Topic modeling. Although we have taken multiple steps to do the cleaning aspects like removing stop words, removing special characters through regular expression, stemming and lemmatization we can do further exploratory data analysis and further clean the dataset for more accurate results.
- Reliability of data available is very important for accurately making inferences of the results of sentiment analysis and Topic modeling. This could be increased by gathering more data about the authenticity of the users which would prevent bias in the results.
- We saw from the data that we have unbalanced rating classes with data biased towards rating 4,5. While predicting the ratings from user reviews it is necessary of balance the data appropriately to remove the bias. This was not incorporated in the current study although we interpreted the results considering the bias. This step could be incorporated to get more accurate results.
- Although we compared 4 models and got a decent accuracy score, we can always venture to new models to gain insights and further enhance our analysis.

References

1. https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17
2. https://towardsdatascience.com/review-rating-prediction-a-combined-approach-538c617c495c https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0
3. https://en.wikipedia.org/wiki/Precision_and_recall
4. http://qpleple.com/topic-coherence-to-evaluate-topic-models/