TEXT MINING ON USER REVIEWS

TO STUDY USER CONCERNS

Group - 3

Abhijith Antony
Hridhay Mehta
Sanjay Jayakumar
Vishnu Vijayakumar

Objective

Exploration of the fields of NLP & Text Mining

Business Recommendation : Employ topic modeling techniques to extract major concerns that the users have about a brand/product (vis-à-vis battery life, design etc.)

Academic Exploration : Predict user ratings based for cell phones on their reviews

Approach

Data sourcing

Data Cleaning & Manipulation

Exploration of the dataset

Sentiment Analysis

Topic Modeling

Predictive Model

University of CINCINNATI | CARL H. LINDNER COLLEGE OF BUSINESS

Data sourcing

Source: Amazon cell-phone reviews dataset from kaggle

items **:**
Pre-scraped data for select 720 phones items

reviews: ~68000 reviews for phone brands in *items*

merged data

	asin	name	rating_x	date	verified	title_x	body	helpfulVotes	brand	title_y	url	image	rating_y	reviewUrl	totalReviews	price	originalPrice
0	B0000SX2UC	Janet	3	October 11, 2005	False	Def not best, but not worst	I had the Samsung A600 for awhile which is abs...	1.0	NaN	Dual-Band / Tri-Mode Sprint PCS Phone w/ Voice...	https://www.amazon.com/Dual-Band-Tri-Mode-Acti...	https://m.media-amazon.com/images/I/2143EBQ210...	3.0	https://www.amazon.com/product-reviews/B0000SX2UC	14	0.0	0.0
1	B0000SX2UC	Luke Wyatt	1	January 7, 2004	False	Text Messaging Doesn't Work	Due to a software issue between Nokia and Spri...	17.0	NaN	Dual-Band / Tri-Mode Sprint PCS Phone w/ Voice...	https://www.amazon.com/Dual-Band-Tri-Mode-Acti...	https://m.media-amazon.com/images/I/2143EBQ210...	3.0	https://www.amazon.com/product-reviews/B0000SX2UC	14	0.0	0.0
2	B0000SX2UC	Brooke	5	December 30, 2003	False	Love This Phone	This is a great, reliable phone. I also purcha...	5.0	NaN	Dual-Band / Tri-Mode Sprint PCS Phone w/ Voice...	https://www.amazon.com/Dual-Band-Tri-Mode-Acti...	https://m.media-amazon.com/images/I/2143EBQ210...	3.0	https://www.amazon.com/product-reviews/B0000SX2UC	14	0.0	0.0

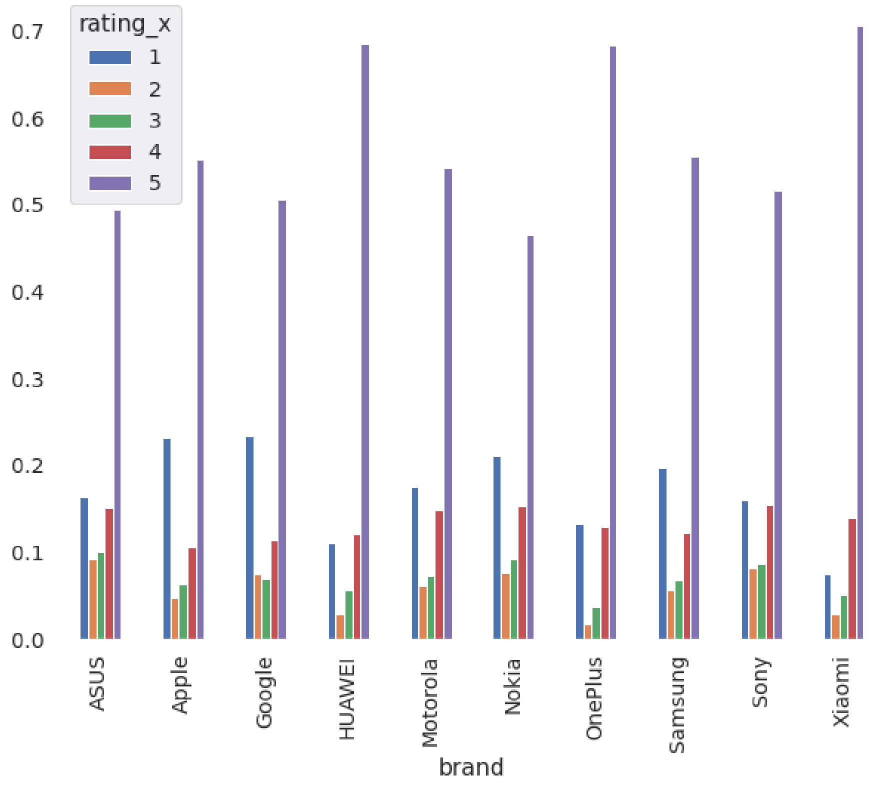
University of CINCINNATI | CARL H. LINDNER COLLEGE OF BUSINESS

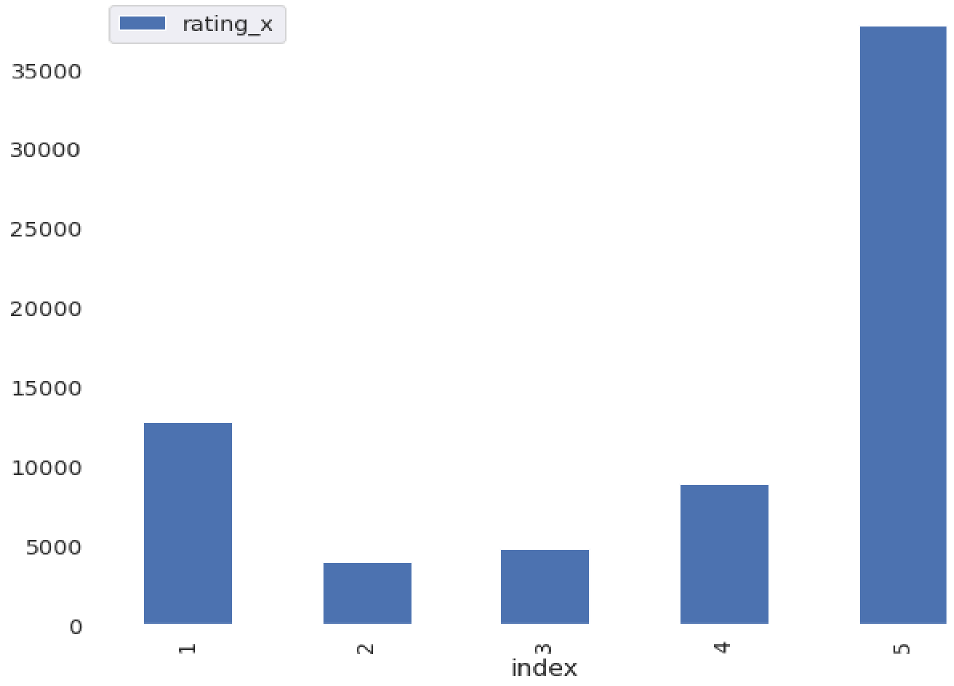
Exploration of the dataset

- We have **67986 reviews** with **17 features**.
- 40% of the reviews are for **Samsung**.
- **Samsung**, **Motorola**, **Nokia** and **Apple** have less than 0.2 rating deviation.

	brand	reviews	avg_avlb_rating	count_phone	avg_rating	total_reviews	data_ratio	rating_dev
0	Samsung	33629	3.781736	346	3.632659	37701	0.891992	0.149077
1	Motorola	8880	3.818694	105	3.643810	9419	0.942775	0.174884
2	Nokia	5915	3.584446	44	3.386364	6182	0.956810	0.198083
3	Apple	5145	3.701263	63	3.782540	6315	0.814727	-0.081276
4	Xiaomi	4411	4.371344	46	4.415217	5574	0.791353	-0.043873
5	Google	3787	3.584896	38	3.771053	4238	0.893582	-0.186157
6	Sony	3196	3.786921	27	3.788889	3312	0.964976	-0.001968
7	HUAWEI	2225	4.240899	32	4.021875	2467	0.901905	0.219024
8	OnePlus	347	4.213256	10	3.580000	406	0.854680	0.633256
9	ASUS	251	3.721116	5	3.860000	263	0.954373	-0.138884

Exploration of the dataset : Distribution of Rating





- ~67% of ratings are 4 & 5 in general
- This is the same for all the brands as well

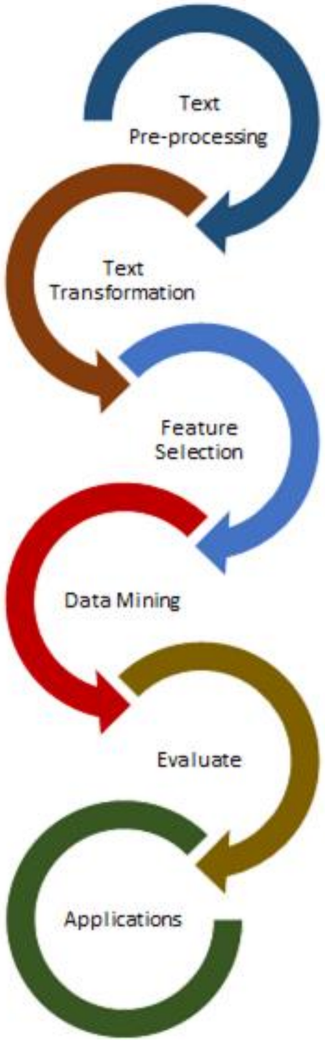
The Companies that we focus on

For this analysis, we focus on 3 major Smartphone companies:

Together, they constitute ~65% of the total smartphone market

University of CINCINNATI | CARL H. LINDNER COLLEGE OF BUSINESS

Text Mining



- Extracting structured information from unstructured text using natural language processing
- High-value knowledge discovery in various areas of application
- R&D, competitive intelligence, patent analysis and market research using sentiment analysis and social media mining
- Algorithms include Entity extraction, Information retrieval, Categorization, Clustering, Summarization

University of CINCINNATI | CARL H. LINDNER COLLEGE OF BUSINESS

Sentiment Analysis

Helps organizations understand the social sentiment of their brand, product or service while monitoring online conversations / reviews

Input : Corpus - A collection of words where order matters.

Output :

- Sentiment score ranging from -1 to 1 (which is polarity)
- Subjectivity score of 0 to 1 where 0 indicates a fact and 1 indicates an opinion.

University of CINCINNATI | CARL H. LINDNER COLLEGE OF BUSINESS

Sentiment analysis

Libraries used

Text Blob

- Python-based open source library
- **Output** : Tuple, polarity, and subjectivity
- Polarity lies within -1.0 & 1.0
- Polarity > 0 = positive
- Polarity < 0 = negative

VADER

- Valence Aware Dictionary & Sentiment
- Protected under MIT license
- **Output** : Polarity score in dictionary format
- Evaluates probability of a positive, negative or neutral sentiment.
- If P(positive) is more, label = positive
- Else, label = negative

University of CINCINNATI | CARL H. LINDNER COLLEGE OF BUSINESS

Sentiment analysis

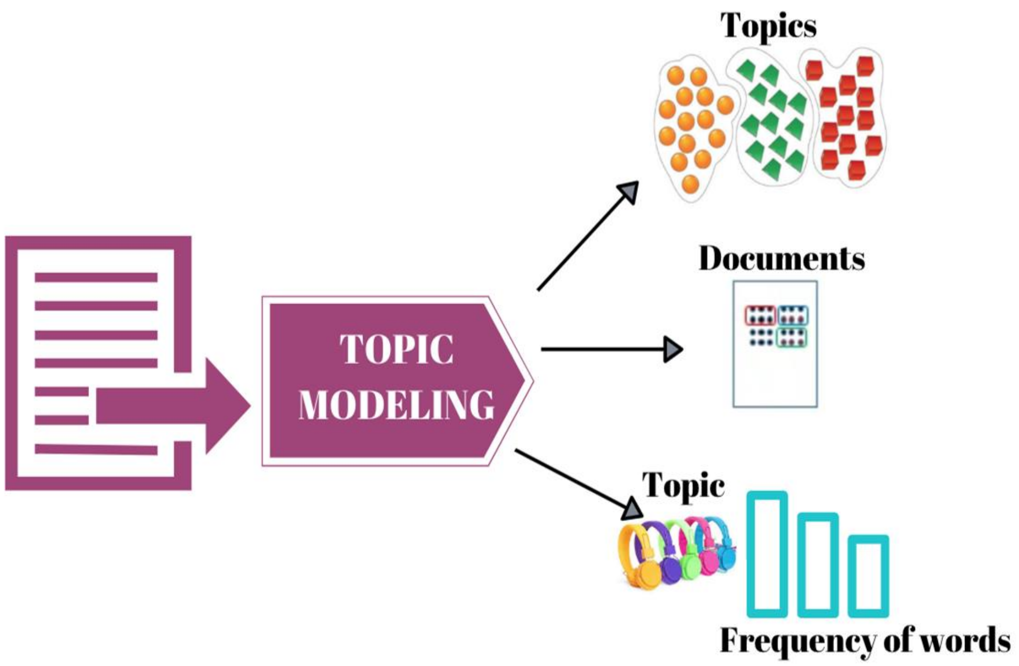
sentiment: negative

rating: 4 or 5

	Unnamed: 0	index	body	brand	rating_x	sent_tb	tb_p	tb_n	tbnb_p	tbnb_n	sent_vader	vader_p	vader_n
24	24	24	SERVED ME WELL AS A BACK UP PHONE.	Motorola	5	0	0	0	0	0	0.2732	0	1
31	31	31	We never use cell phones, but thought we neede...	Motorola	5	0	0	0	0	0	-0.6526	0	1
37	37	37	I almost never write reviews for anything, but...	Motorola	5	0	0	0	0	0	-0.9063	0	1
43	43	43	This phone isn't kidding when it says military...	Motorola	4	0	0	0	0	0	-0.8005	0	1
45	45	45	i wont (enter special code)	Motorola	5	0	0	0	0	0	-0.3089	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
67948	67948	67948	Sorry, this video is unsupported on this brows...	Sony	5	0	0	0	0	0	0.4372	0	1
67950	67950	67950	En general es uno de los mejores Xperia que he...	Sony	5	0	0	0	0	0	0.0000	0	1
67953	67953	67953	Quility	Sony	5	0	0	0	0	0	0.0000	0	1
67968	67968	67968	forget about iPhones. I've been using xperia s...	Sony	5	0	0	0	0	0	0.2398	0	1
67983	67983	67983	buy one more for my cousin	Sony	5	0	0	0	0	0	0.0000	0	1

	brand	avg_sent_tb	tb_p	tb_n	avg_sent_vader	vader_p	vader_n
0	ASUS	0	0	0	0.450343	177	74
1	Apple	0	0	0	0.303738	3124	2021
2	Google	0	0	0	0.371158	2549	1238
3	HUAWEI	0	0	0	0.456779	1499	726
4	Motorola	0	0	0	0.405286	6120	2760
5	Nokia	0	0	0	0.373347	4018	1897
6	OnePlus	0	0	0	0.488430	244	103
7	Samsung	0	0	0	0.361681	22199	11430
8	Sony	0	0	0	0.447593	2263	933
9	Xiaomi	0	0	0	0.396646	2592	1819

Topic Modeling



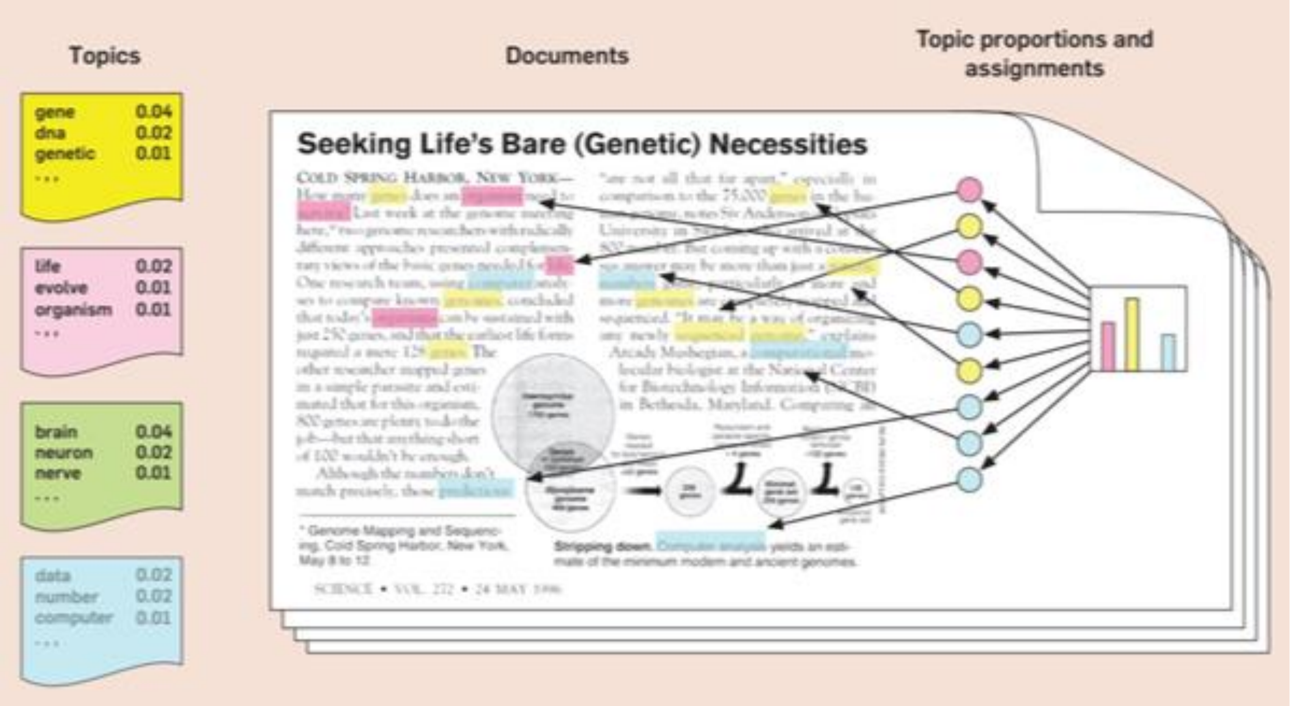
- Discovers abstract “topics” in a collection of documents

- Automatically “learns” groups or clusters of words that best characterize the data

- Major algorithms
  - [Latent Dirichlet Allocation]
  - Structural Topic Models
  - [Non-negative matrix factorization]

Linear Dirichlet Allocation

Concept : Each document can be described by a distribution of topics
Each topic can be described by a distribution of words



Visualizing the findings of LDA Model

- Our analysis is focussed on these 3 companies : Samsung, Apple and Xiaomi.

- The subsequent slides will show the following 2 aspects on a brand-wise level

  - Word cloud of top 10 words in each topic

  - The most discussed topics in the documents
    - By weight in that document
    - By summing up the weight contribution of each topic to the document

University of CINCINNATI | CARL H. LINDNER COLLEGE OF BUSINESS

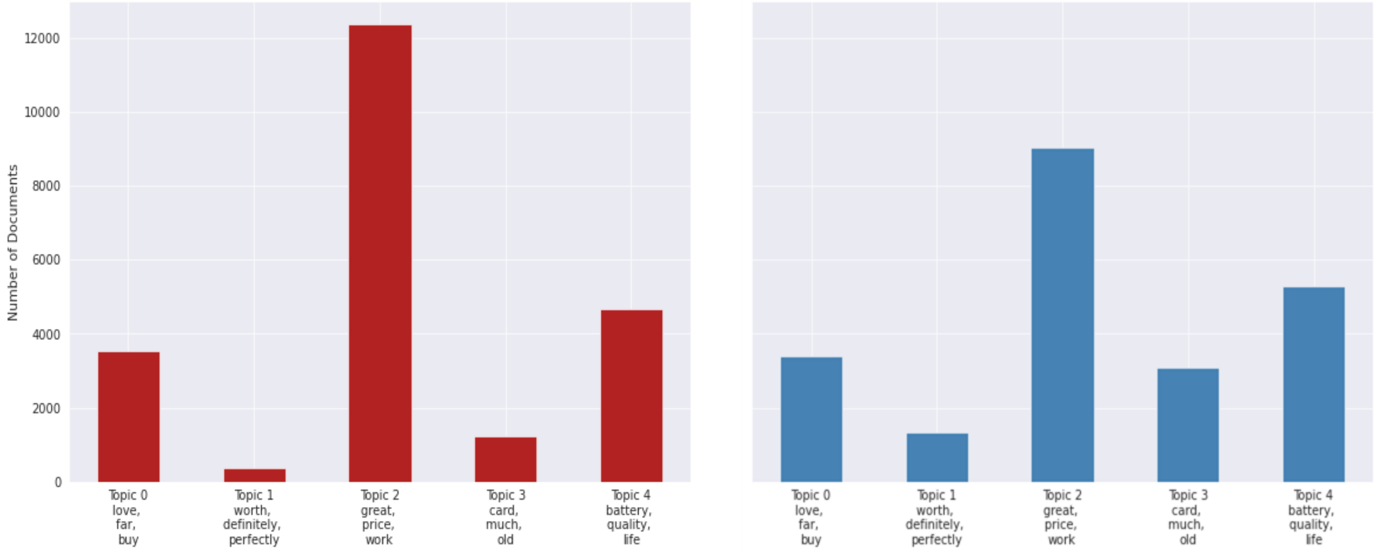
Positive comments

Word cloud

Most Discussed Topics

by dominant topic

by topic weightage



Takeaways:

1. Users tend to love their purchase experience
2. User find the product worth the money spent
3. Users like the quality & durability of product

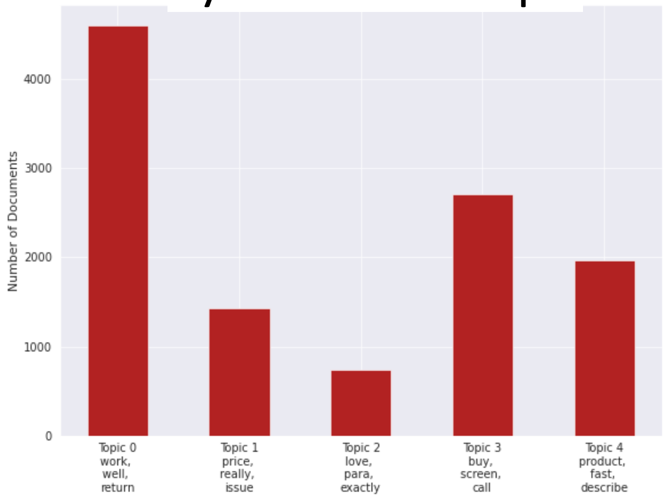
University of CINCINNATI | CARL H. LINDNER COLLEGE OF BUSINESS

Negative comments

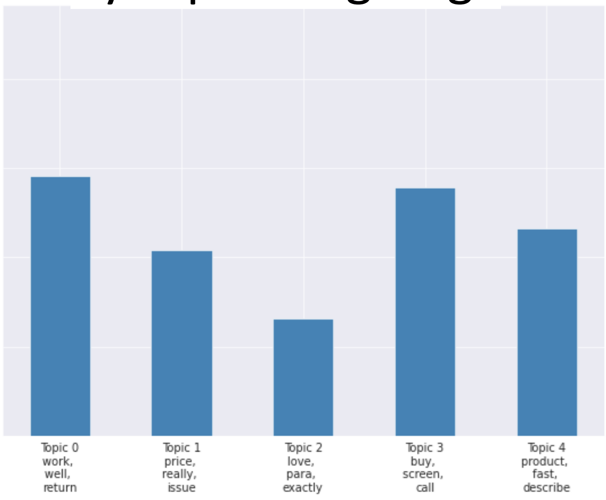
Word cloud

Most Discussed Topics

by dominant topic



by topic weightage



Takeaways:

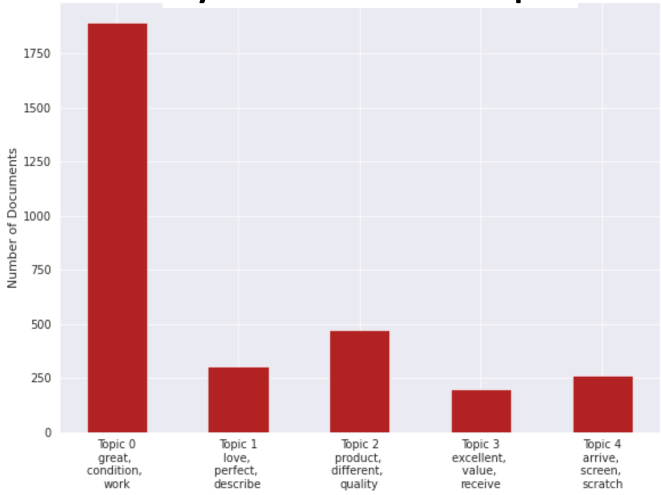
1. Users have an issue with the working of product such as heating issue
2. Users have problems with camera , screen etc
3. Users do not find the product to be fast

iPhone Positive comments

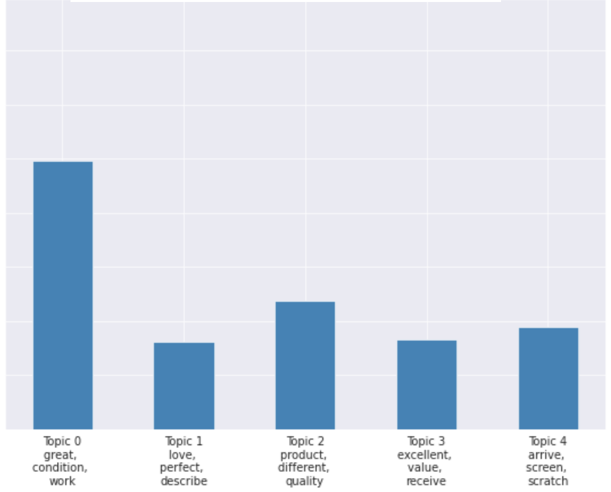
Word cloud

Most Discussed Topics

by dominant topic



by topic weightage

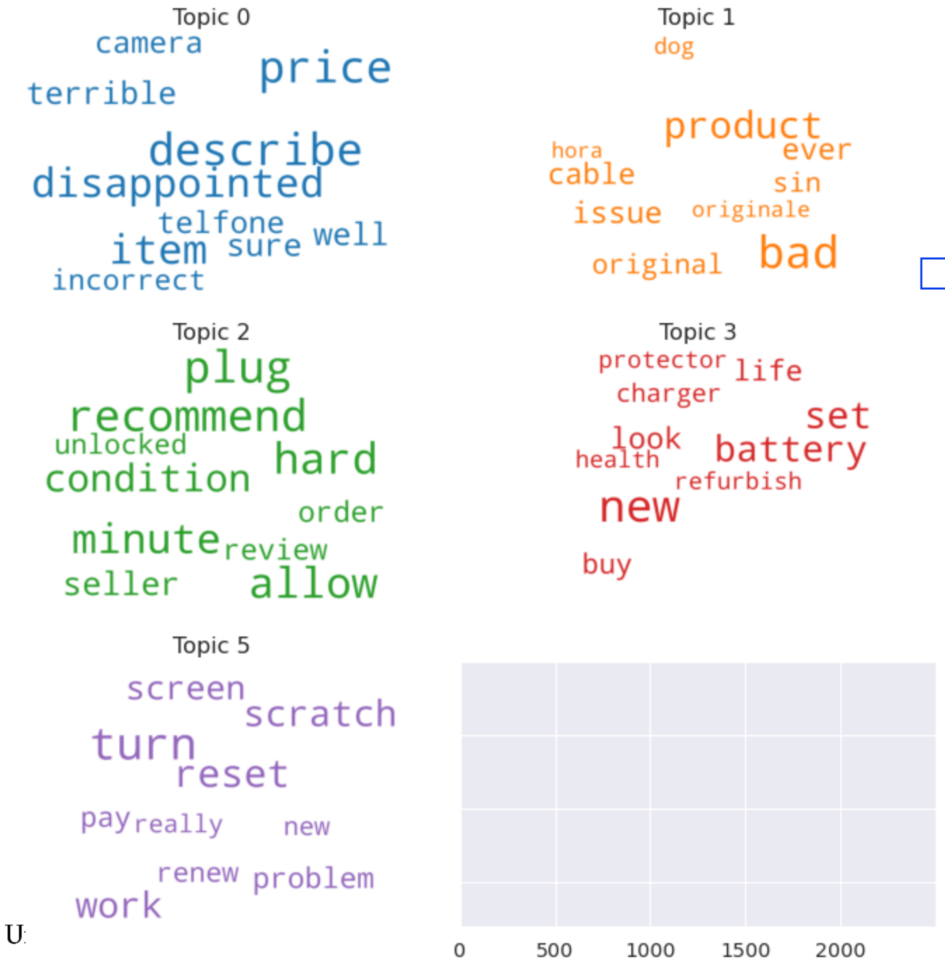


Takeaways:

1. Users find product to be great buy
2. Users are positive about the quality
3. Users like the seller qualities

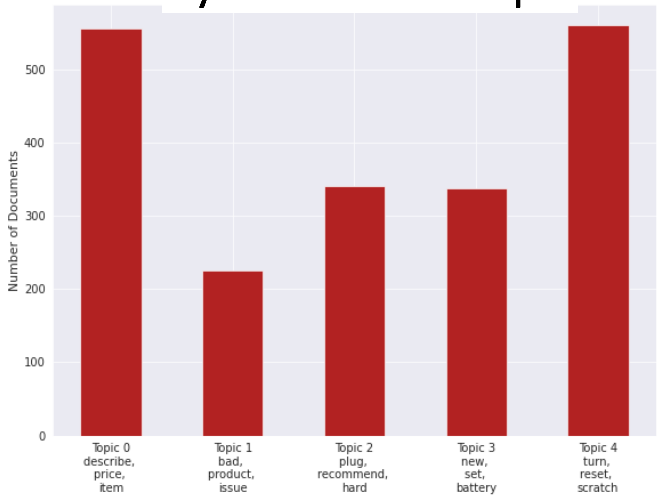
iPhone Negative comments

Word cloud

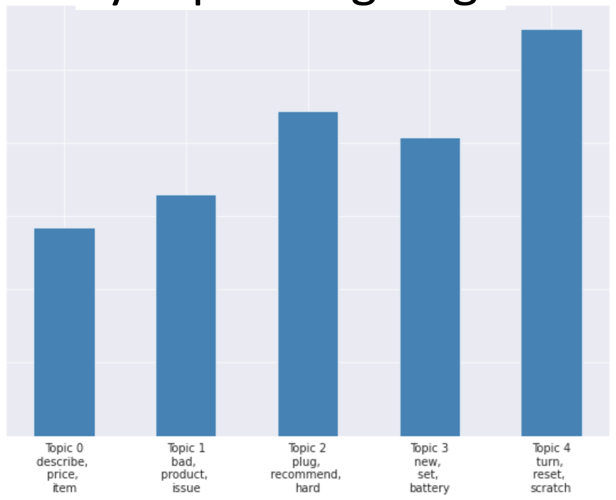


Most Discussed Topics

by dominant topic



by topic weightage



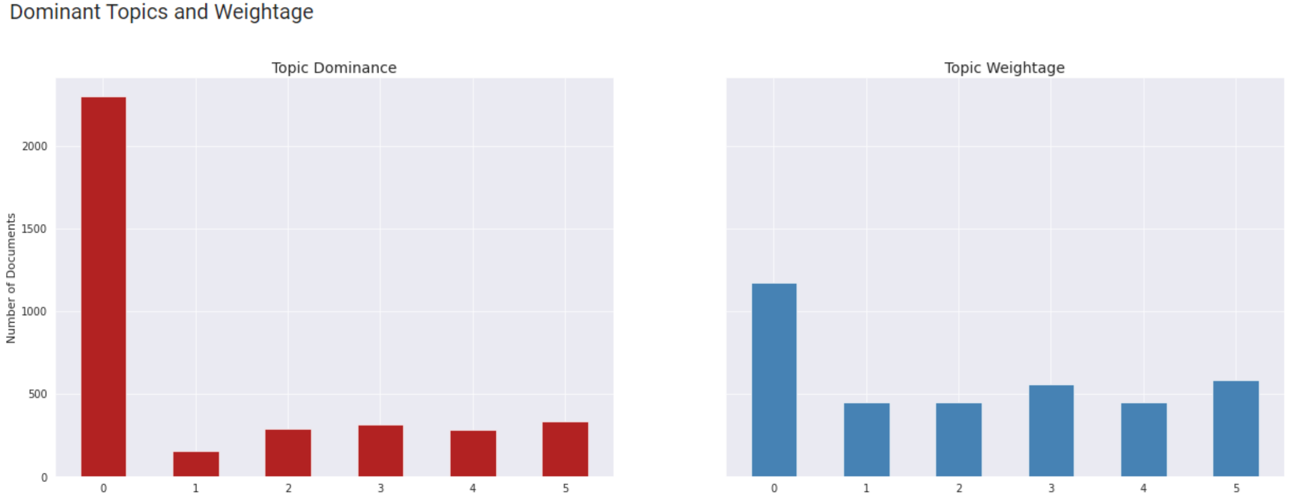
Takeaways:

1. Users are divided on the priciness of product
2. Product screen has scratches on arrival
3. Users have issues with the plug

Xiaomi Positive comments

Word cloud

Most Discussed Topics



Takeaways:

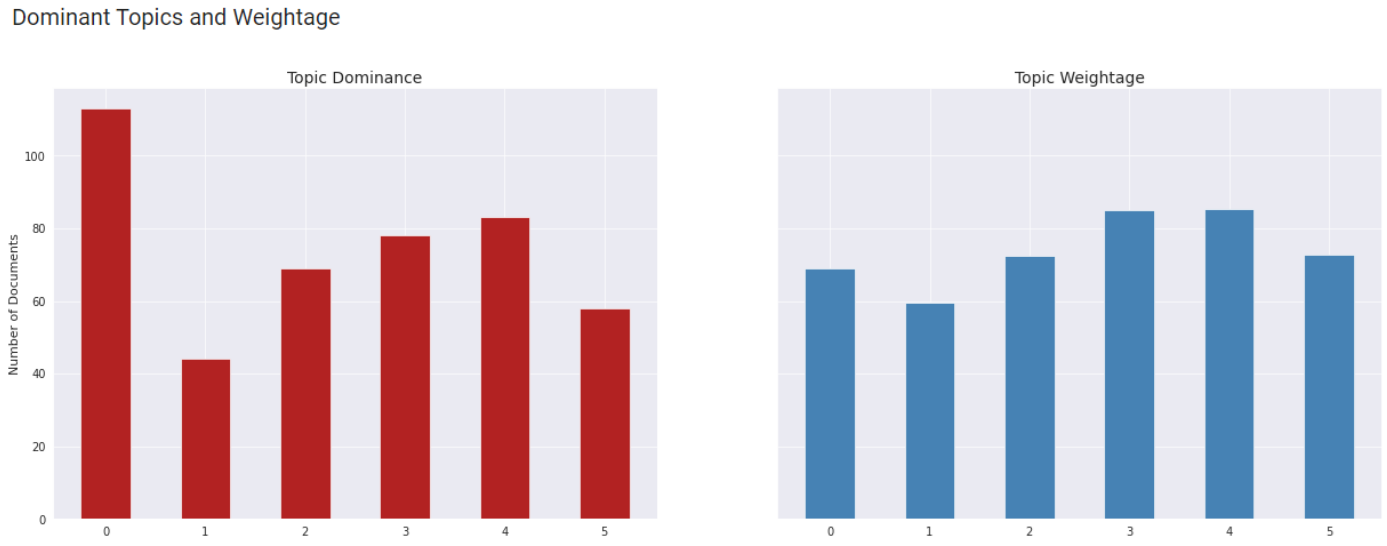
1. Users find the product to be a great value for money
2. Users like the camera and battery aspects of the product

University of CINCINNATI | CARL H. LINDNER COLLEGE OF BUSINESS

Xiaomi **Negative comments**

Word cloud

Most Discussed Topics



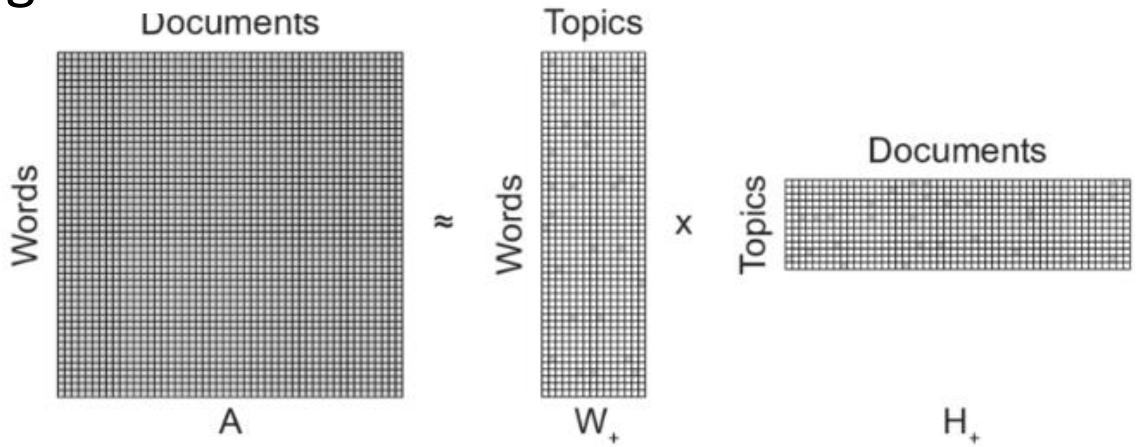
Takeaways:

1. Users expect a premium product
2. Users have issues with the charger and storage

University of CINCINNATI | CARL H. LINDNER COLLEGE OF BUSINESS

Non Matrix Factorisation

- Using linear algebra for topic modeling (in essence)
- Input: Term-Document matrix, number of topics.
- Output: Two non-negative matrices of the original n words by k topics and those same k topics by the m original documents.



Conceptual illustration of non-negative matrix factorization (NMF) decomposition of a matrix consisting of m words in n documents into two non-negative matrices of the original n words by k topics and those same k topics by the m original documents.

University of CINCINNATI | CARL H. LINDNER COLLEGE OF BUSINESS

Results of NMF model : Xiaomi

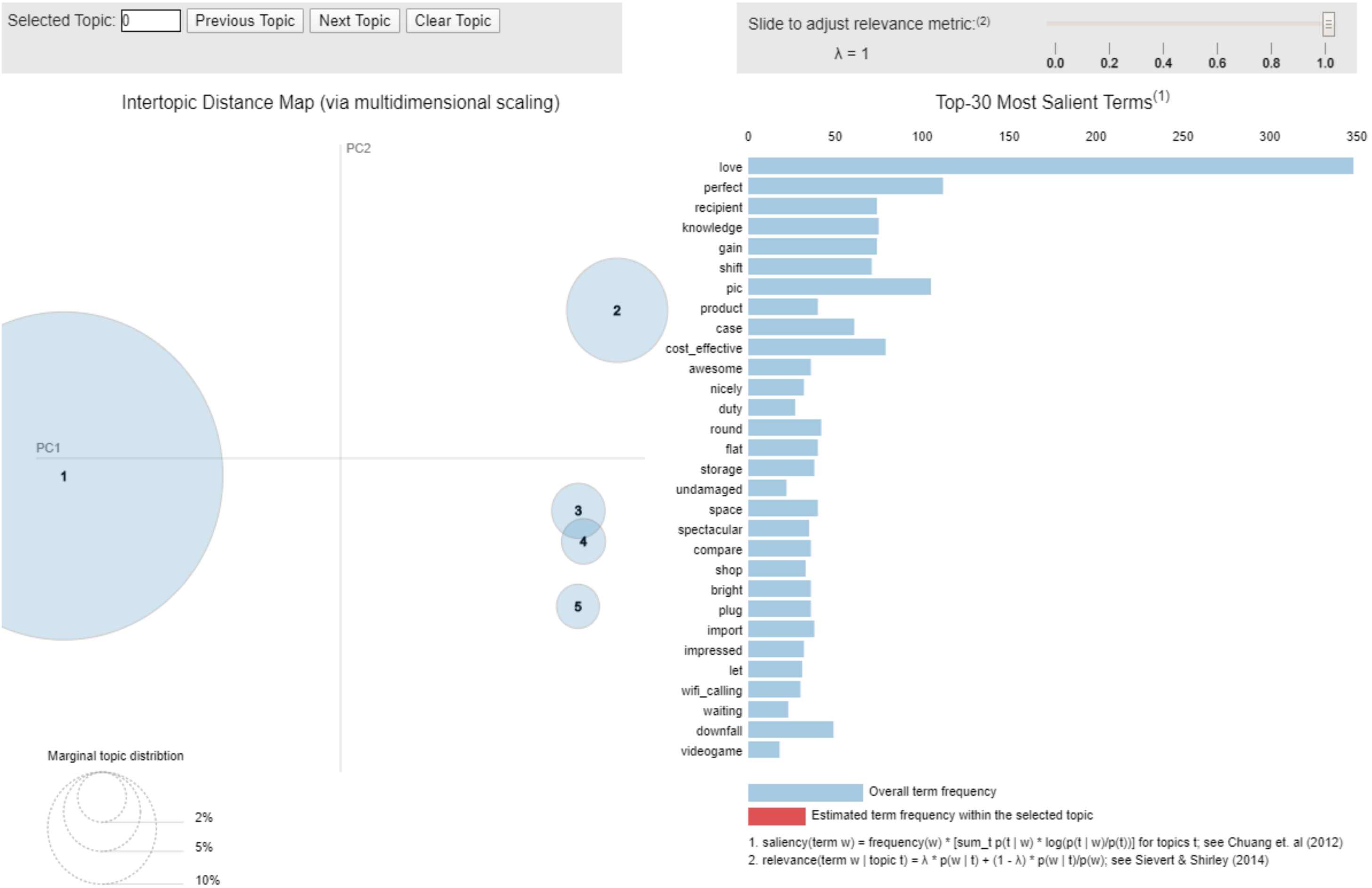
	Topic # 01	Topic # 02	Topic # 03	Topic # 04	Topic # 05	Topic # 06
0	great price	value money	excellent product	best ever	amazing price	work great
1	price buy	great value	product perfect	ever buy	absolutely amazing	amazing work
2	great great price	great value money	incomparable service	life day	really 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
11	price wish	battery well	product money	give best	thing redemi	great well
12	price great	value money price	camera overall	happy purchase	price compete	great picture
13	battery life	exceed expectation	money pay	love much	price cost	great product
14	buy love	money price	sun long	work great camera	worth pay	money work
15	son love	way better	sun long operate	better old	price compare	worth money
16	mobile lte	international charger	display look great	much work	price still	name brand
17	price excellent	price performance	display look	great battery life	great amazing	brand work
18	hard find	incredible value	small hand	great battery	thing consider	great fast
19	price hard	great battery	fast smooth	picture video	pretty price	great far

	Topic # 01	Topic # 02	Topic # 03	Topic # 04	Topic # 05	Topic # 06
0	sorry video	rede mejore	basicamente compre	stop charge	sin embargo	picture quality
1	unsupported browser	sluggish nature	month battery	charge day	sin audio	value disappoint
2	video unsupported browser	sin sin	tengo nuevo	battery bad	sin audio sin	rear camera picture
3	sorry video unsupported	fast charge	varias ocasione	uninstalled send	audio sin	value disappoint rear
4	video unsupported	pero inferior	battery low	user issue	audio sin embargo	rear camera
5	unsupported browser charge	sin errore	qualification camera	return date	month battery	month battery
6	short cell	compatible metro_pc	utilice publicacion	problem find user	sluggish nature	work great fix
7	update short	turn ad hence	sin parede	return date problem	varias ocasione	rear camera work
8	update short cell	turn ad	show issue	unfortunately return date	pero inferior	worth return
9	unsupported browser buy	tough cheap option	show issue charge	unfortunately return	utilice publicacion	send back
10	buy day	tough cheap	purchase look spec	problem find	battery low	camera work
11	unfortunately screen return	top level	purchase life	side camera ok	tengo nuevo	work great
12	unfortunately screen	screen fast	purchase look	replace wife	qualification camera	sin errore
13	sad exited unfortunately	week freak	weak battery	side camera	sin sin	posible sin
14	unsupported browser super	screen fast processor	bad purchase	replace wife screen	purchase look spec	price unbeatable
15	sad exited	screen grab	charge battery	really sharp fast	show issue charge	relative price
16	return back	screen grab replace	relative price	sharp fast	purchase life	sin parede
17	manufacturer defect	screen money	receive call datum	sharp fast really	purchase look	advertisement everywhere
18	refund percentage restocking	screen money break	purchase due support	really milliamp	show issue	unlocking process
19	quite time seller	option hit mark	slowly surely	really sharp	bad purchase	wifi terrible

University of CINCINNATI | CARL H. LINDNER COLLEGE OF BUSINESS

Interactive visualisation from Gensim package

[Link to the Google Colab file]()



University of CINCINNATI | CARL H. LINDNER COLLEGE OF BUSINESS

Predicting user ratings from their reviews

- Models used : Logistic regression, Naïve Bayes, SVM, Random Forest models.

- Steps followed:
  - Convert text data into TF-IDF vectors
  - Split the data into a training and test set
  - Classify the text data using different models

- Evaluate the performance of each of the models using precision, recall and accuracy.

Predicting user ratings from their reviews

Metrics related to the models mentioned earlier are as given:

	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**
	Naive 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**