Finding a Helpful Phone Review

Group5

Cara Groden, Chih-Jung Tu,
Henrique Peretti, Hsin-Yu Hsieh

Business Case: Using NLP to Improve Sales and Customer Experience

01 Business Insights from Customers

We'd like to understand our customers better through the reviews they leave. What's important to them and what are they looking for?

02 Show Customers Similar / Relevant Reviews by Others

To improve our website, we'd like to suggest to customers products and reviews that match what they're looking for.

For example, if a customer wants to find a phone with a great camera that is easy to operate, they could search for “a phone with a great camera and easy to operate” and our site would find relevant reviews or topics based on similar mentions of the camera and/or ease of use.

Data Source and Data Cleaning

Data Source

A 127mb dataset of more than 400,000 Amazon's review about unlocked mobile phone sales.

Variables	
Product Name	String
Brand Name	String
Price	Float (0.00 to infinite)
Rating (1)	Integer (1 to 5)
Reviews	String
Review Votes (2)	Float

Data Cleaning

Initial dataset - 413,840 observations

Remove Blanks

Remove blank reviews and reviews with missing votes values

Replace Characters

Remove embedded html characters
Remove punctuation
Convert to lower case
Remove extra white space

Cleaned dataset - 401,482 observations

(1) Rating assigned by the user in the review
(2) Number of people who found the review helpful

Implementation Approach

01 **Deduplication**

Remove duplicates of reviews using MinHash and Spark.

02 **Topic Modeling**

Use LDA model to discover the different topics that the reviews represent and top words for each topic.

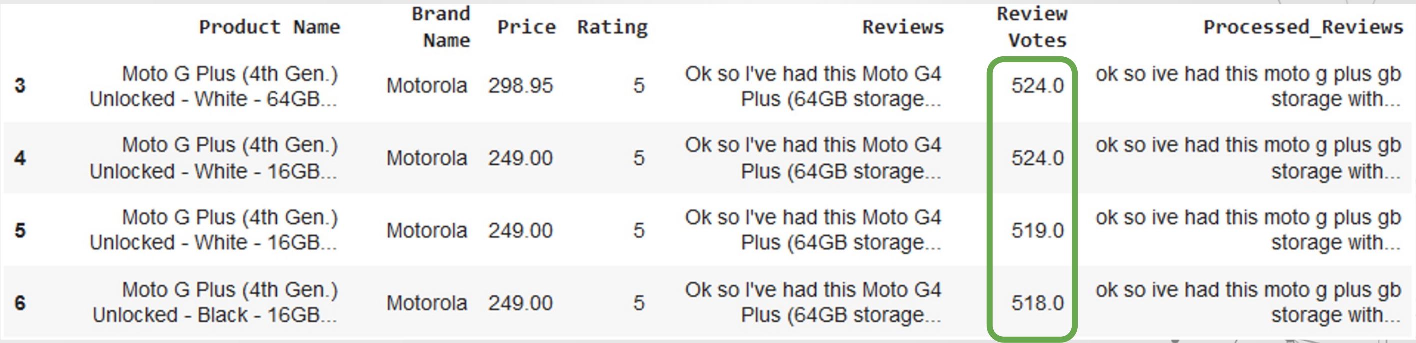
03 **Finding Similarity**

Train a Word2Vec model to measure the semantic similarity of reviews and input terms to show reviews that are most similar to a customer’s search input.

Deduplication - Implementation

- Original plan to use SimHash and Word2Vec changed as implementation using Spark and MinHash was more successful.
  - Jaccard distance, strings of variable length.
- Caveat: Spark is very slow to run over a dataset this size and we had to run it with just a subset of our data.

- Sorted reviews by “helpful” votes - deduplication would identify most “helpful” review
- Note same review text associated with product variations (storage size and phone color, etc.) and votes for same review can vary.



	Product Name	Brand Name	Price	Rating	Reviews	Review Votes	Processed_Reviews
3	Moto G Plus (4th Gen.) Unlocked - White - 64GB...	Motorola	298.95	5	Ok so I've had this Moto G4 Plus (64GB storage...	524.0	ok so ive had this moto g plus gb storage with...
4	Moto G Plus (4th Gen.) Unlocked - White - 16GB...	Motorola	249.00	5	Ok so I've had this Moto G4 Plus (64GB storage...	524.0	ok so ive had this moto g plus gb storage with...
5	Moto G Plus (4th Gen.) Unlocked - White - 16GB...	Motorola	249.00	5	Ok so I've had this Moto G4 Plus (64GB storage...	519.0	ok so ive had this moto g plus gb storage with...
6	Moto G Plus (4th Gen.) Unlocked - Black - 16GB...	Motorola	249.00	5	Ok so I've had this Moto G4 Plus (64GB storage...	518.0	ok so ive had this moto g plus gb storage with...

Deduplication - Evaluation

When run on a small sample of our data (5,000 rows), we were able to very accurately identify duplicate reviews.

	Subset of Dataset	Full Dataset (cleaned)
Rows Considered (Cleaned Dataset)	**5,000**	**401,482**
Duplicates Found using Deduplication	2,103	N/A
Actual Unique Reviews Found using: `len(set(review_text))`	**2,103**	**156,021**
Duplicates as % of Dataset	42.2%	38.7%

Topic Modeling - LDA

LDA is a generative probabilistic model that assumes each topic is a mixture over an underlying set of words, and each document is a mixture of over a set of topic probabilities.

Process :
given the M number of documents, N number of words, and prior K number of topics, the model trains to output:

1. the distribution of words for each topic K
2. the distribution of topics for each document i

LDA Implementation

1. Prepare data for LDA
   - ➢ Tokenize data
   - ➢ Remove Stopwords
   - ➢ Convert to vectors

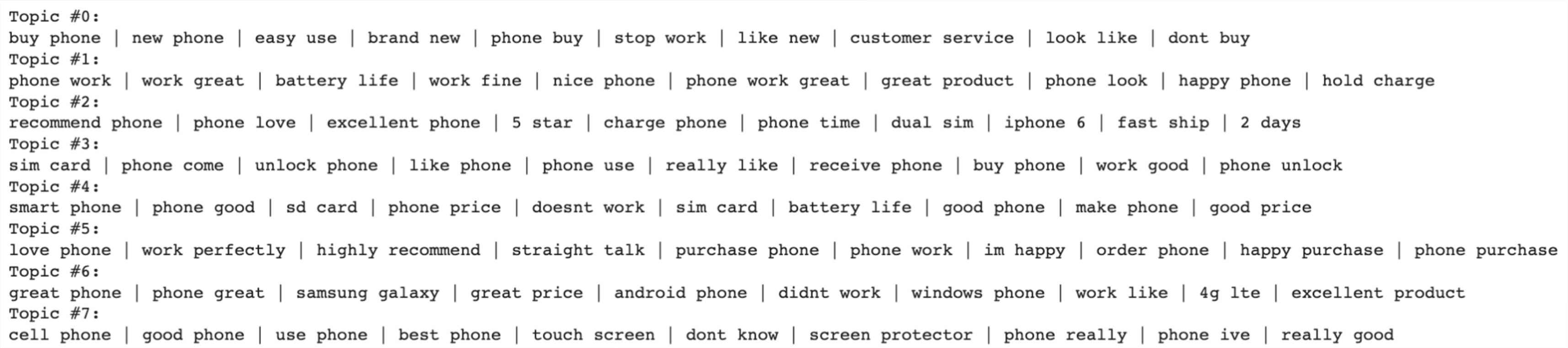
1. LDA model training
   - ➢ SKLearn LDA

1. Coherence Score

Topic Modeling - Findings

8 topics and 10 top words for each topic

LDA model results over cleaned data



```
Topic #0:
buy phone | new phone | easy use | brand new | phone buy | stop work | like new | customer service | look like | dont buy
Topic #1:
phone work | work great | battery life | work fine | nice phone | phone work great | great product | phone look | happy phone | hold charge
Topic #2:
recommend phone | phone love | excellent phone | 5 star | charge phone | phone time | dual sim | iphone 6 | fast ship | 2 days
Topic #3:
sim card | phone come | unlock phone | like phone | phone use | really like | receive phone | buy phone | work good | phone unlock
Topic #4:
smart phone | phone good | sd card | phone price | doesnt work | sim card | battery life | good phone | make phone | good price
Topic #5:
love phone | work perfectly | highly recommend | straight talk | purchase phone | phone work | im happy | order phone | happy purchase | phone purchase
Topic #6:
great phone | phone great | samsung galaxy | great price | android phone | didnt work | windows phone | work like | 4g lte | excellent product
Topic #7:
cell phone | good phone | use phone | best phone | touch screen | dont know | screen protector | phone really | phone ive | really good
```

Topics of customer reviews that are directly related to products:

phone condition, workability, battery, ease of use, sim card, sd card, screen, appearance, price

Topics that are not directly related to products:

shipping, customer service

Other topics:

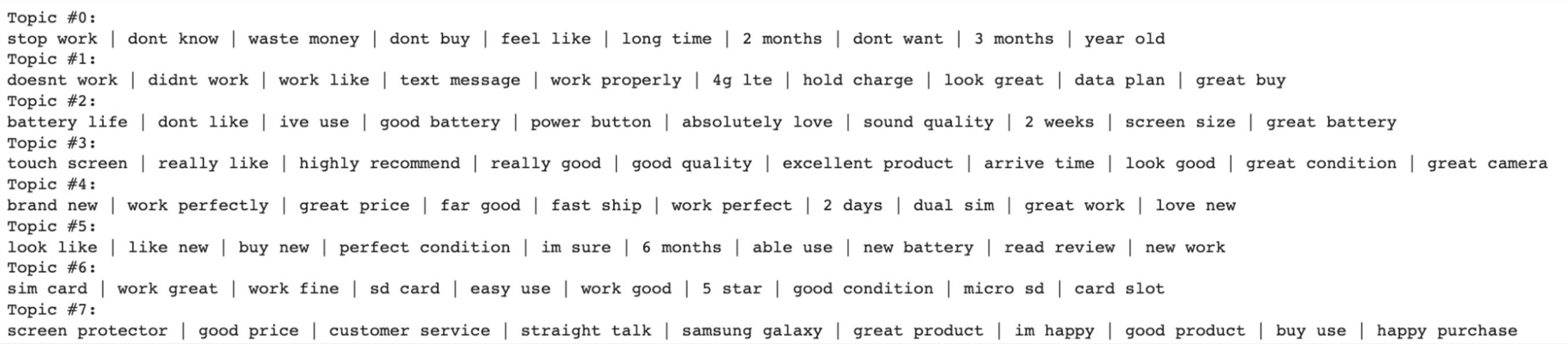
customers' feelings towards the purchase, whether customers themselves would recommend the product

Overall, it shows that customers care about not only the product itself but also shipping, service, and the overall experience.

Topic Modeling - Findings

8 topics and 10 top words for each topic

LDA model results over cleaned data removing ‘phone’



```
Topic #0:
stop work | dont know | waste money | dont buy | feel like | long time | 2 months | dont want | 3 months | year old
Topic #1:
doesnt work | didnt work | work like | text message | work properly | 4g lte | hold charge | look great | data plan | great buy
Topic #2:
battery life | dont like | ive use | good battery | power button | absolutely love | sound quality | 2 weeks | screen size | great battery
Topic #3:
touch screen | really like | highly recommend | really good | good quality | excellent product | arrive time | look good | great condition | great camera
Topic #4:
brand new | work perfectly | great price | far good | fast ship | work perfect | 2 days | dual sim | great work | love new
Topic #5:
look like | like new | buy new | perfect condition | im sure | 6 months | able use | new battery | read review | new work
Topic #6:
sim card | work great | work fine | sd card | easy use | work good | 5 star | good condition | micro sd | card slot
Topic #7:
screen protector | good price | customer service | straight talk | samsung galaxy | great product | im happy | good product | buy use | happy purchase
```

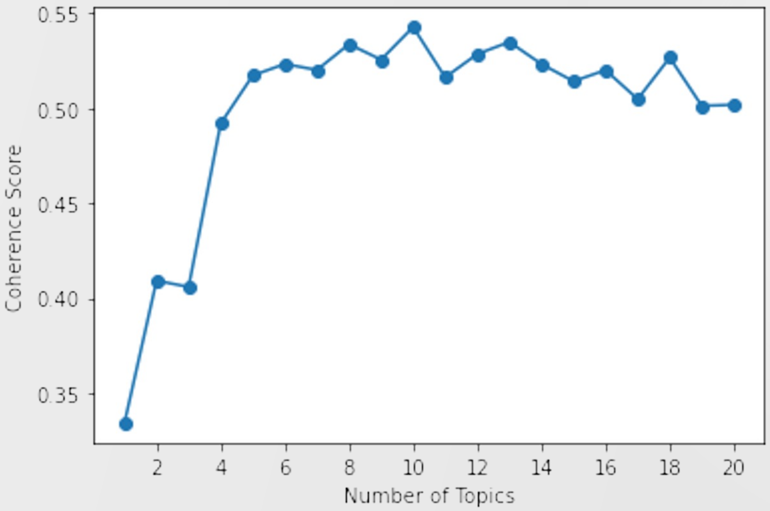
Specific topics after removing ‘phone’ in the data:

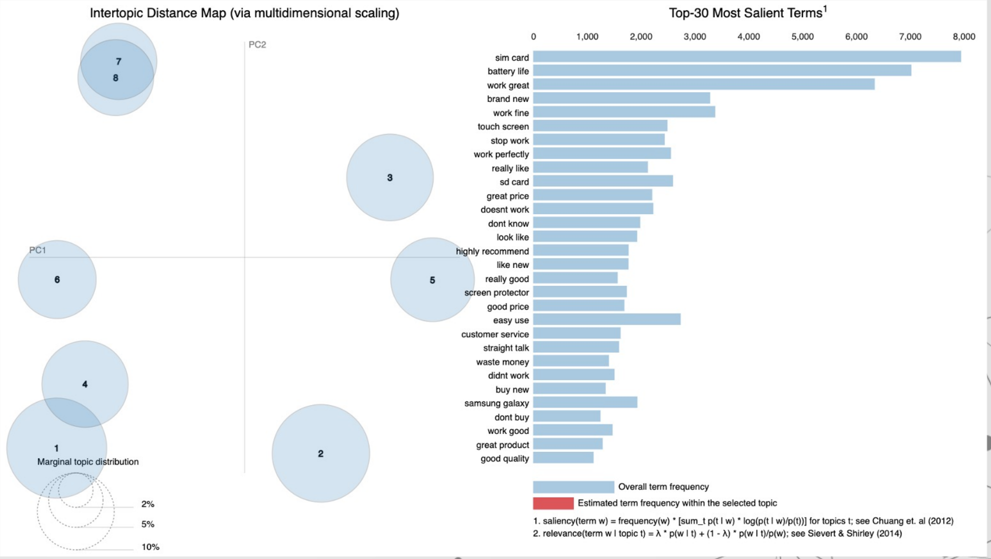
data plan, power button, sound quality, camera

The results still cover topics like phone condition, price, and battery life. But after removing ‘phone’ in the data, LDA model results reveal more detailed information about products. For example, when talking about sim cards and sd cards, details like dual sim, micro sd, and card slot appear in the results. The information can be helpful for product developers, sellers, and many more.

LDA with Topic Coherence

Use topic coherence to measure the degree of semantic similarity between high scoring words in the topic. We can use the coherence score in topic modeling (the “elbow” method) to ensure we have a meaningful number of topic groups.





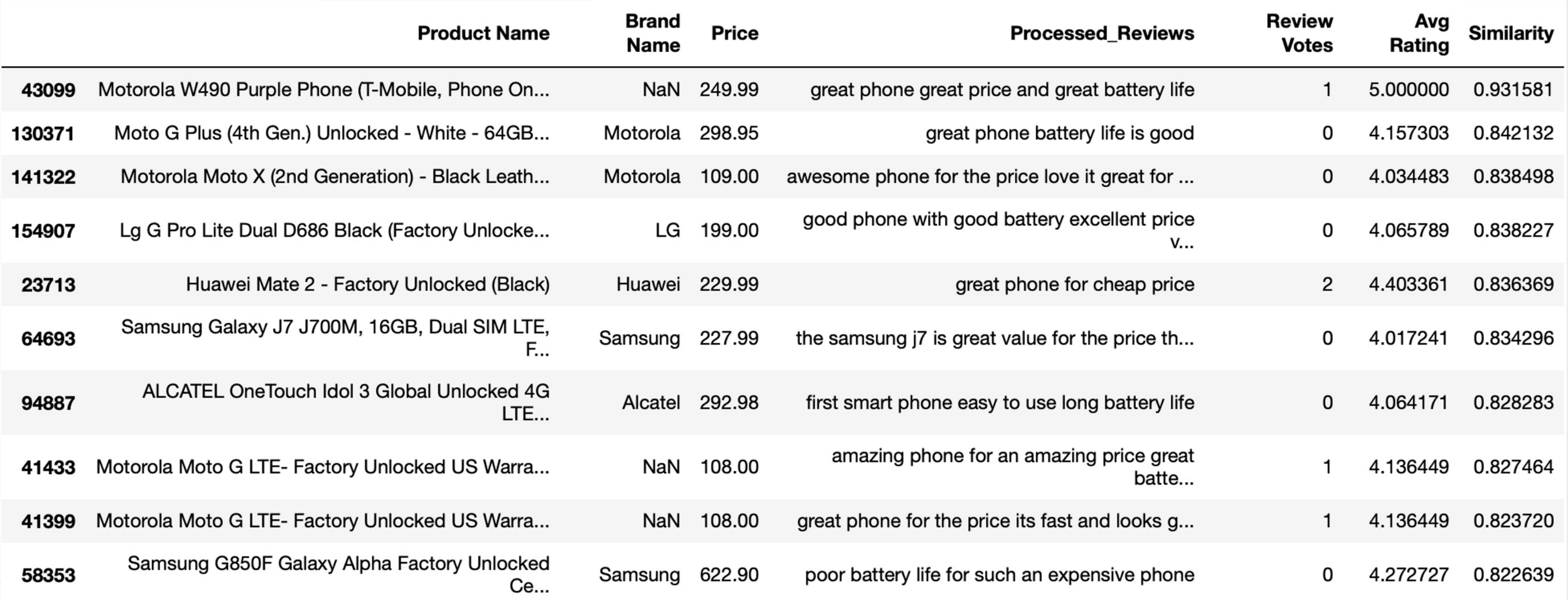
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Finding Semantic Similarity

- We aim to find similar reviews to a given search input by the customer. The customer can input a list of words, phone features and adjectives and we will show phones with **a review containing similar words, disregarding whether that is the most appropriate product**
- We created a column called "Similarity" to store the similarity index for each comparison.
- We calculate the similarity between two strings (Amazon review and customer text) using a Word Vector Model. Word2Vec groups the vector of similar words together in the vector space. The similarity is an index ranging from 0 to 1; the higher the "closer" spatially or more similar they are
- We use the clean and deduplicated dataset to run the similarity calculation. Comparing clean texts proved to be computationally faster
- **Limitations of the search:** this is a semantic similarity between review text and input search, the objective is not to find the best product for the given words.

Similarity Results

Query "great phone cheap price long battery life"



	Product Name	Brand Name	Price	Processed_Reviews	Review Votes	Avg Rating	Similarity
43099	Motorola W490 Purple Phone (T-Mobile, Phone On...	NaN	249.99	great phone great price and great battery life	1	5.000000	0.931581
130371	Moto G Plus (4th Gen.) Unlocked - White - 64GB...	Motorola	298.95	great phone battery life is good	0	4.157303	0.842132
141322	Motorola Moto X (2nd Generation) - Black Leath...	Motorola	109.00	awesome phone for the price love it great for ...	0	4.034483	0.838498
154907	Lg G Pro Lite Dual D686 Black (Factory Unlocke...	LG	199.00	good phone with good battery excellent price v...	0	4.065789	0.838227
23713	Huawei Mate 2 - Factory Unlocked (Black)	Huawei	229.99	great phone for cheap price	2	4.403361	0.836369
64693	Samsung Galaxy J7 J700M, 16GB, Dual SIM LTE, F...	Samsung	227.99	the samsung j7 is great value for the price th...	0	4.017241	0.834296
94887	ALCATEL OneTouch Idol 3 Global Unlocked 4G LTE...	Alcatel	292.98	first smart phone easy to use long battery life	0	4.064171	0.828283
41433	Motorola Moto G LTE- Factory Unlocked US Warra...	NaN	108.00	amazing phone for an amazing price great batte...	1	4.136449	0.827464
41399	Motorola Moto G LTE- Factory Unlocked US Warra...	NaN	108.00	great phone for the price its fast and looks g...	1	4.136449	0.823720
58353	Samsung G850F Galaxy Alpha Factory Unlocked Ce...	Samsung	622.90	poor battery life for such an expensive phone	0	4.272727	0.822639

In order to solve the limitations, we are only showing the top 10 similar results with products with average rating above 4.0

Conclusions

Challenges

1. The biggest challenge seems to be running the code over the whole dataset. It might be more feasible with more computing resources, which would be an investment for any company trying to implement the project.
2. The reviews are sometimes associated with the wrong products so it would be challenging to iterate further using price or product as criteria.
3. Similarity index is imperfect, some results were misleading: for example "long battery life" and "not great battery life" were very similar. Additionally, some reviews were in different languages and therefore will generate a low similarity index albeit could be very similar in concept.

4. The similarity search has limitations, the objective is to show similar reviews for well rated products (average rating above 4), but not necessarily the results will reflect the best possible products..

Improvement

1. When sorting by similarity we lose the best reviews (i.e. the ones with more helpful votes). We could have an option of calculating similarity for most helpful reviews only.