

Opinion Mining: Information Extraction for Product Development

Ayan Karim

Business Problem

- ▶ Any company that sells a product in mass production generates a huge amount of public opinion on the product and tons of reviews, articles, reactions and overall sentiments are posted online, especially soon after release.
- ▶ As of 2017, 223 million iPhones were sold in the US, which is huge considering the population was 325 million (this information was from 2017)!
- ▶ How do companies access this data from the public media and turn it into something useful? How do we extract all that diverse information from the web and analyze it??

Introduction

- ▶ What is Opinion Mining?
- ▶ What is Aspect-Based Sentiment Analysis

Let's Pretend We're Apple!



What does Public Opinion tell us about the iPhone X?

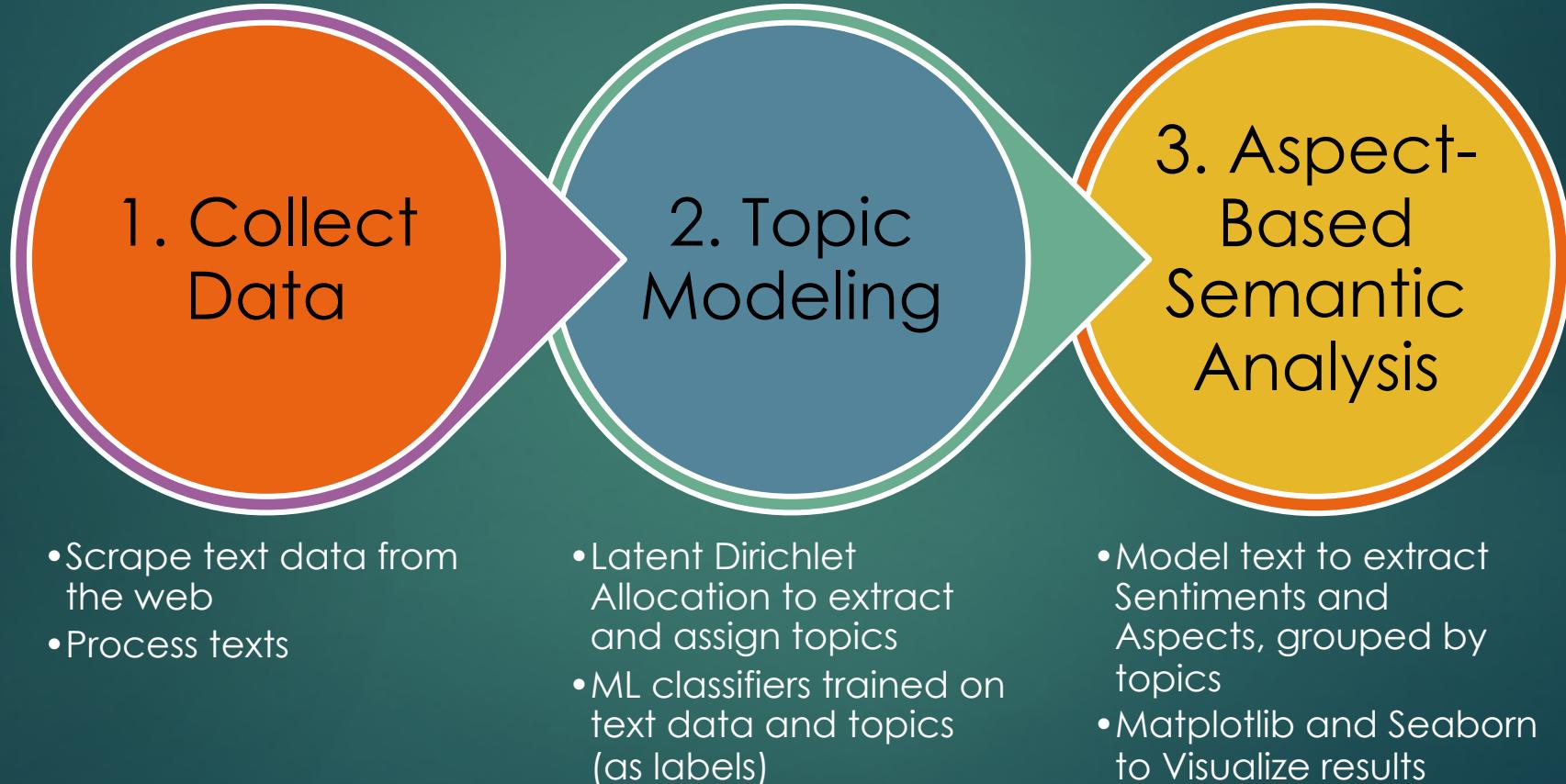


What do people dislike about the iPhone X?



How does our product compare with that of a competing smartphone like the Samsung Galaxy S9?

Pipeline



1. 2. Collecting Data and Topic Modeling

digitaltrends.com
gizmodo.com
techradar.com

- ▶ 1. iPhone X
- ▶ 2. Galaxy S9
- ▶ 3. Pixel 3
- ▶ 4. Huawei Mate 20 Pro
- ▶ 5. OnePlus 6T
- ▶ 6. Huawei P2 Pro
- ▶ 7. LG V4 ThinQ
- ▶ 8. Sony Xperia XZ3
- ▶ 9. Essential Phone
- ▶ 10. Razer Phone 2
- ▶ 11. HTC U12+
- ▶ 12. Moto G6 plus



Shrinking demand forces Apple to slow down iPhone X production

The iPhone X was a major hit when it launched last holiday season, but demand has slowed since then. Various industry sources report that Apple is expected to produce as few as eight million units in the second quarter of 2018.

Posted 4.28.18 — By Eric Brackett



Worth it. The iPhone X is the breath of fresh air Apple fans were waiting for

★★★★★

After 10 years of evolution, the Apple iPhone X pushes design forward with a noticeably different design, even if the guts resemble the more modest iPhone 8 and 8 Plus. Here's our review.

Posted 5.14.18 — By Julian Chokkattu

	author	text	title
0	Julian Chokkattu	Google's annual hardware launch event will tak...	Google will announce hardware on October 9, ne...
1	Christian de Looper	Google finally unveiled the new Google Pixel 3...	Here's how to buy the new Google Pixel 3 and G...
2	Simon Hill	If you plan to buy one of Google's Pixel 3 sma...	The best Pixel 3 cases and covers
3	Simon Hill	As the developer of Android, Google turns out ...	Google Pixel 3 vs. Pixel 2 vs. Pixel: Picking ...
4	Simon Hill	There are plenty of contenders in the Android ...	Google Pixel 3 vs. Samsung Galaxy S9: Which sm...
5	Julian Chokkattu	Got your hands on a new Pixel 3 or Pixel 3 XL ...	Key settings you need to change on your brand...
6	Simon Hill	Rarely has a flagship phone been so thoroughly...	Google Pixel 3 and Pixel 3 XL: Everything you ...
7	Lucas Coll	Mobile hardware is getting better and better, ...	Verizon's buy one, get one offer is the best d...
8	Christian de Looper	The Google Pixel 3 and Pixel 3 XL may have sto...	The Google Pixel Stand turns your Android phon...
9	Simon Hill	The Google Pixel 3 and Pixel 3 XL are phones w...	The best Google Pixel 3 tips and tricks

Out[20]:

Selected Topic: 1

Previous Topic

Next Topic

Clear Topic

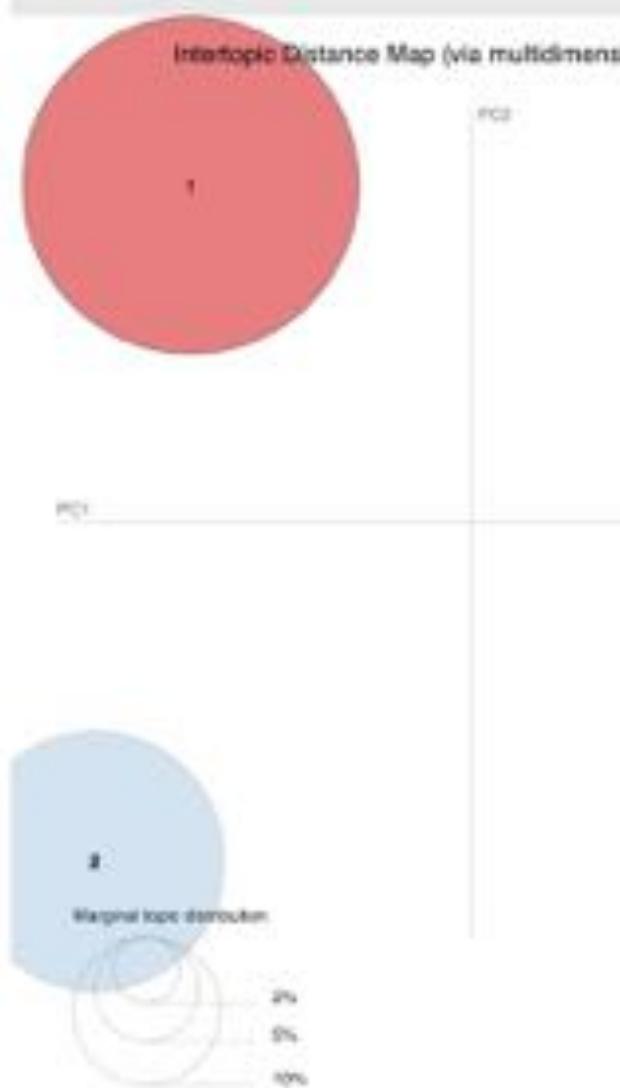
Slide to adjust relevance matrix (0)

3 = 1

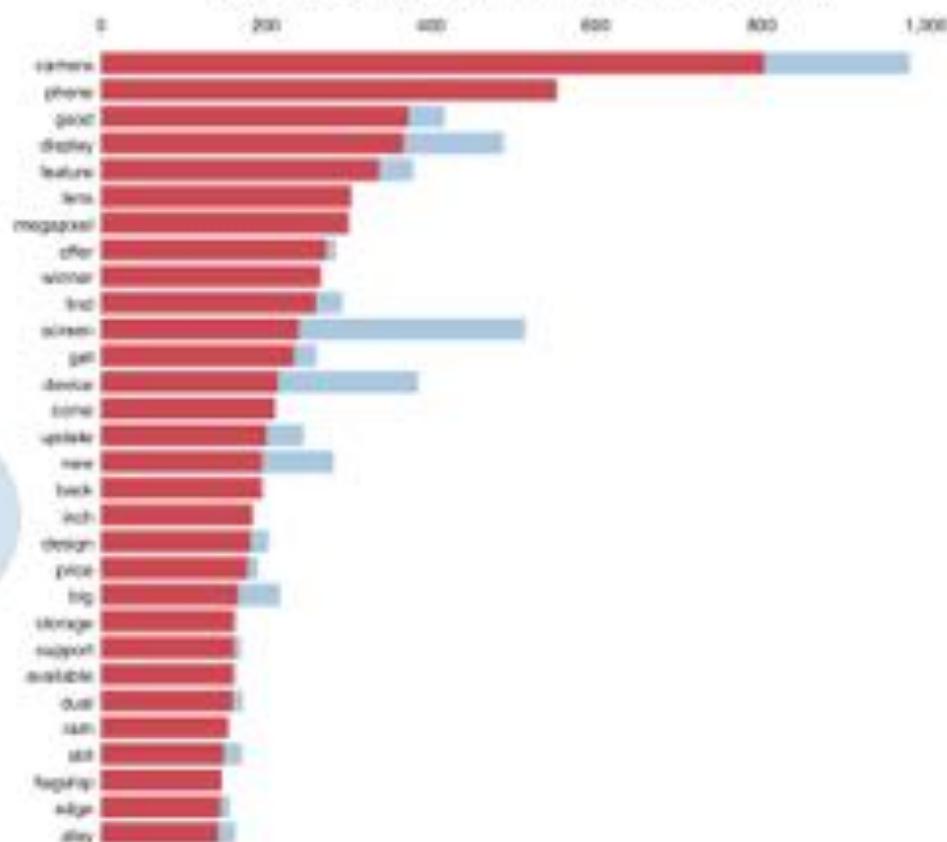
0.0 0.2 0.4 0.6 0.8



Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 1 (52% of tokens)



Overall term frequency

Estimated term frequency within the selected topic

1. Author's internal validation (from 1 to 100) based on expert knowledge for topics 1–5; see Chauhan et al. (2012).

2. Author's internal validation (from 1 to 100) based on expert knowledge for topics 6–10; see Chauhan & Shrivastava (2013).

```
[ (0,
  '0.018*"camera" + 0.013*"phone" + 0.009*"good" + 0.008*"display" +
  '0.008*"feature" + 0.007*"lens" + 0.007*"megapixel" + 0.006*"offer" +
  '0.006*"winner" + 0.006*"find" + 0.006*"screen" + 0.005*"get" +
  '0.005*"device" + 0.005*"come" + 0.005*"update"""),
(1,
  '0.018*"app" + 0.017*"screen" + 0.014*"case" + 0.011*"tap" + 0.008*"home" +
  '0.007*"set" + 0.006*"want" + 0.006*"display" + 0.006*"setting" +
  '0.006*"option" + 0.006*"go" + 0.006*"turn" + 0.006*"button" + 0.005*"use" +
  '0.005*"time"""),
(2,
  '0.007*"photo" + 0.006*"even" + 0.006*"camera" + 0.006*"device" +
  '0.005*"could" + 0.004*"say" + 0.004*"take" + 0.004*"user" + 0.004*"issue" +
  '0.004*"really" + 0.004*"thing" + 0.004*"leak" + 0.003*"people" +
  '0.003*"new" + 0.003*"year" )]
```

Latent Dirichlet Allocation

Labels



DESIGN



FUNCTION



RELIABILITY



	author	text	title	labelled
0	Julian Chokkattu	Google's annual hardware launch event will tak...	Google will announce hardware on October 9, ne...	[design]
1	Christian de Looper	Google finally unveiled the new Google Pixel 3...	Here's how to buy the new Google Pixel 3 and G...	[reliability]
2	Simon Hill	If you plan to buy one of Google's Pixel 3 sma...	The best Pixel 3 cases and covers	[function]
3	Simon Hill	As the developer of Android, Google turns out ...	Google Pixel 3 vs. Pixel 2 vs. Pixel: Picking ...	[reliability]
4	Simon Hill	There are plenty of contenders in the Android ...	Google Pixel 3 vs. Samsung Galaxy S9: Which sm...	[reliability]

Assign Labels

```
In [28]: # LabelPowerset allows for multi-label classification
# Build a pipeline for multinomial naive bayes classification
text_clf = Pipeline([('vect', CountVectorizer(stop_words = "english", norm='l1', binary=True)),
                     ('tfidf', TfidfTransformer(use_idf=False)),
                     ('clf', LabelPowerset(MultinomialNB(alpha=0.001), n_jobs=-1))])

text_clf = text_clf.fit(X_train, y_train)
predicted = text_clf.predict(X_test)

# Calculate accuracy
np.mean(predicted == y_test)

Out[28]: 0.809256254254254
```



```
In [29]: # Test if SVM perform better
text_clf_svm = Pipeline([('vect', CountVectorizer()),
                        ('tfidf', TfidfTransformer()),
                        ('clf-svm', LabelPowerset(
                            SGDClassifier(loss='hinge', penalty='l2',
                                          alpha=0.0001, max_iter=10, random_state=42))))])

text_clf_svm = text_clf_svm.fit(X_train, y_train)
predicted_svm = text_clf_svm.predict(X_test)

#Calculate accuracy
np.mean(predicted_svm == y_test)

/usr/local/lib/python3.6/site-packages/sklearn/linear_model/stochastic_gradient.py:183: FutureWarning: max_iter and tol parameters have been added first in 0.19. If max_iter is set but tol is left unset, the default value for tol is 0.19 and 0.20 will be None (which is equivalent to -infinity, effect) but will change in 0.21 to 1e-3. Specify tol to silence this warning.
  FutureWarning)

Out[29]: 0.8564814814814815
```

Multi-Label Support Vector Machine Classifier

3. Aspect-Based Sentiment Analysis

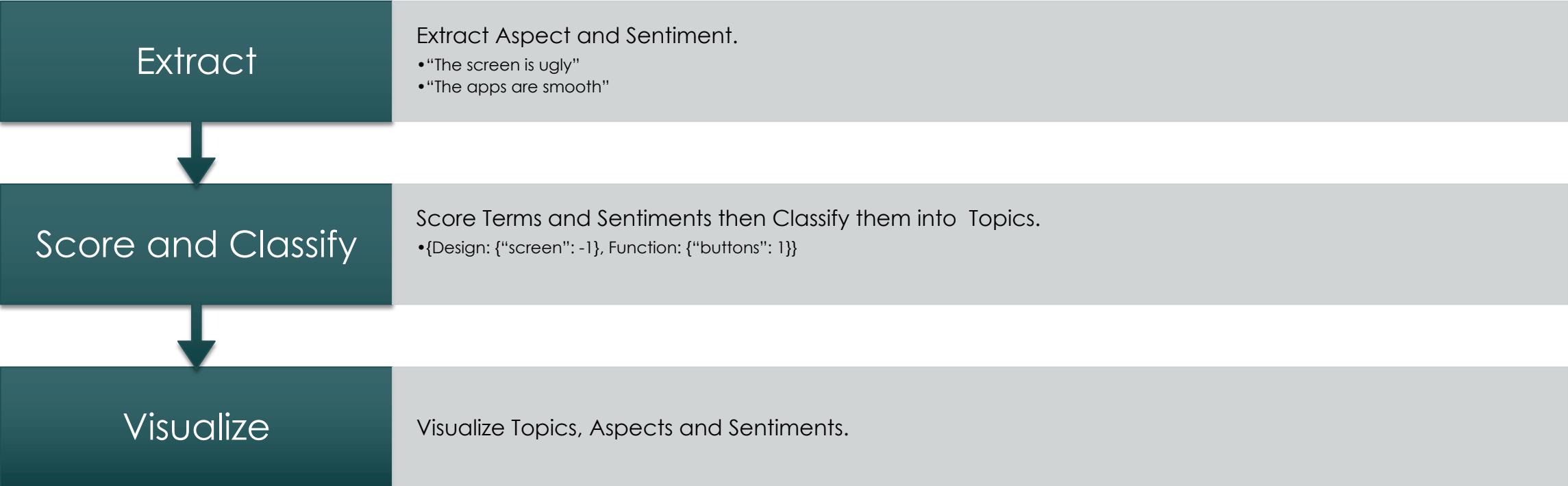
digitaltrends.com
gizmodo.com
techradar.com



1. IPHONE X



2. GALAXY S9



Aspect-Based Sent. Analysis Pipeline

Visualize Topics, Aspects and Sentiments.

Negative Aspects of iPhone X

design	function	reliability
('fall', -2)	('shame', -2)	('limit', -2)
('cracks', -2)	('things', -2)	('dust', -2)
('hack', -2)	('trouble', -2)	('anti', -2)
('marketing', -2)	('zoom', -2)	('belief', -2)
('handset', -2)	('reception', -2)	('number', -2)
('lens', -2)	('swipe', -2)	('cost', -3)
('bugs', -3)	('blur', -2)	('problems', -4)
('bumps', -4)	('falls', -3)	('limits', -4)
('scratch', -6)	('shock', -3)	('complaints', -5)
('scratches', -8)	('damage', -4.5)	('noise', -6)

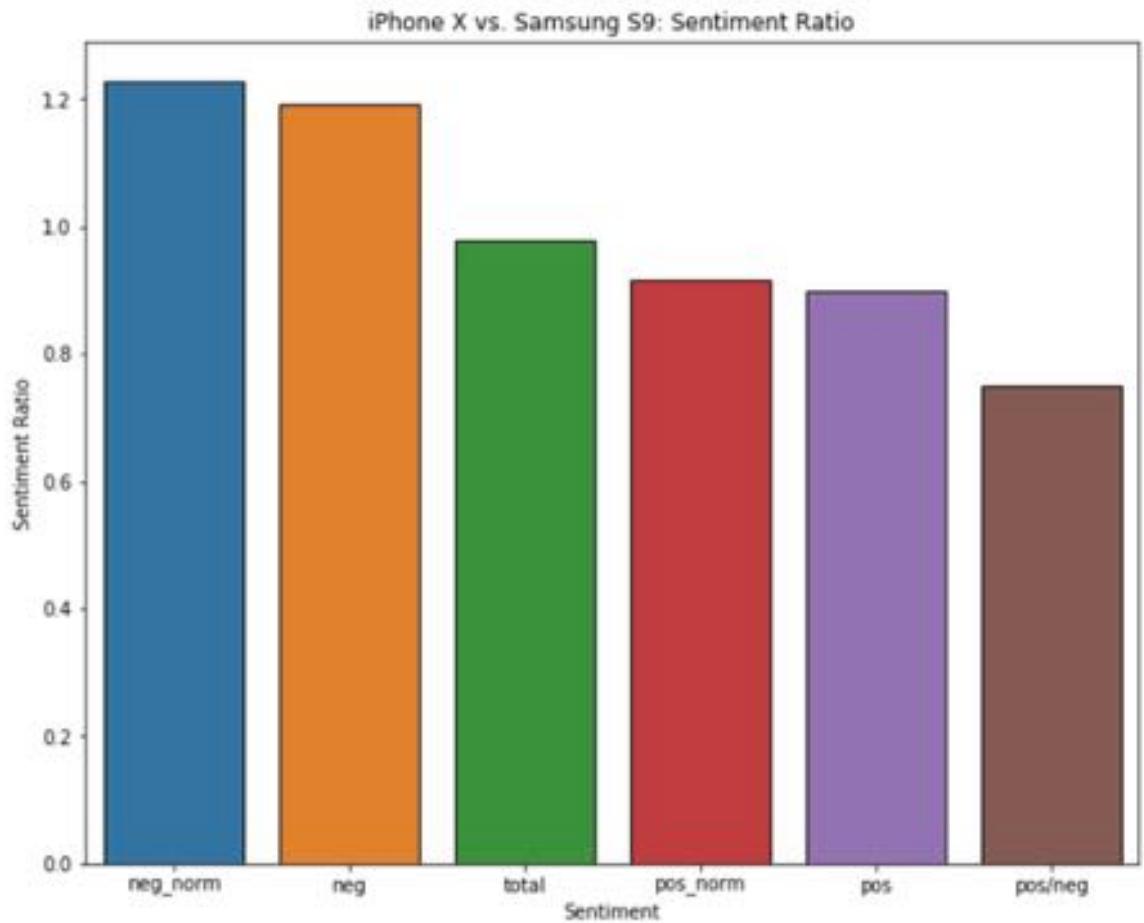
Negative Aspects of Samsung Galaxy S9

design	function	reliability
('attention', -1)	('punch', -2)	('dark', -1)
('sales', -2)	('touch', -2)	('vibration', -1.5)
('picture', -2)	('tricks', -2)	('downgrade', -1.5)
('lens', -2)	('stroke', -2)	('error', -2)
('bugs', -2)	('gimmick', -2)	('complaints', -2)
('shader', -3)	('finger', -2)	('limits', -3)
('browser', -3)	('words', -2)	('battery', -3.5)
('fall', -3)	('lag', -2)	('leaks', -5)
('flaw', -5)	('shame', -4)	('noise', -5.5)
('video', -6)	('motion', -10)	('lack', -6.5)

	iPhoneX design	iPhoneX function	iPhoneX reliability	S9 design	S9 function	S9 reliability	iPhoneX vs. Samsung S9 (ratio)
neg	134.500000	130.500000	121.500000	108.000000	108.000000	108.000000	1.192901
pos	308.000000	225.500000	243.000000	288.500000	288.500000	288.500000	0.897169
total	442.500000	356.000000	364.500000	396.500000	396.500000	396.500000	0.977722
neg_norm	0.303955	0.366573	0.333333	0.272383	0.272383	0.272383	1.228491
pos_norm	0.696045	0.633427	0.666667	0.727617	0.727617	0.727617	0.914465
pos/neg	2.289963	1.727969	2.000000	2.671296	2.671296	2.671296	0.750938

iPhone X vs. Samsung Galaxy S9

iPhone X vs. Samsung Galaxy S9



Next Steps

- ▶ 1. Set up a feature for standardized streaming of data that updates our datasets in intervals.
- ▶ 2. Develop a recommendation engine to recommend which aspects should be most actionable for product development based on our results.

Improvements

- ▶ 1. Collect more data
- ▶ 2. Combine synonyms as the same aspect, so that there aren't duplicate information
- ▶ 3. The Algorithms have trouble interpreting some vocabulary like bigrams and trigrams, as well as very tech specific terminology, we need to make this vocab interpretable somehow.

Bibliography

- ▶ "Apple iPhone sales 2018." Statista. Statista. 19 Feb. 2019 <<https://www.statista.com/statistics/263401/global-apple-iphone-sales-since-3rd-quarter-2007/>>.
- ▶ Bansal, Shivam, and Natural Language Processing and Machine Learning. "Beginners Guide to Topic Modeling in Python." Analytics Vidhya. 11 Jan. 2019. 19 Feb. 2019 <<https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/>>.
- ▶ Li, Susan. "Topic Modeling and Latent Dirichlet Allocation (LDA) in Python." Towards Data Science. 31 May 2018. Towards Data Science. 19 Feb. 2019 <<https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24>>.
- ▶ Li, Susan. "Topic Modelling in Python with NLTK and Gensim – Towards Data Science." Towards Data Science. 30 Mar. 2018. Towards Data Science. 19 Feb. 2019 <<https://towardsdatascience.com/topic-modelling-in-python-with-nltk-and-gensim-4ef03213cd21>>.
- ▶ Min, Peter. "Aspect-Based Opinion Mining (NLP with Python) – Peter Min – Medium." Medium.com. 06 June 2018. Medium. 19 Feb. 2019 <<https://medium.com/@pmin91/aspect-based-opinion-mining-nlp-with-python-a53eb4752800>>.
- ▶ <https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>
- ▶ <http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/>
- ▶ <https://www.digitaltrends.com/>
- ▶ <https://www.gizmodo.com/>
- ▶ <https://www.techradar.com/>