Analyzing Customer Support in Twitter

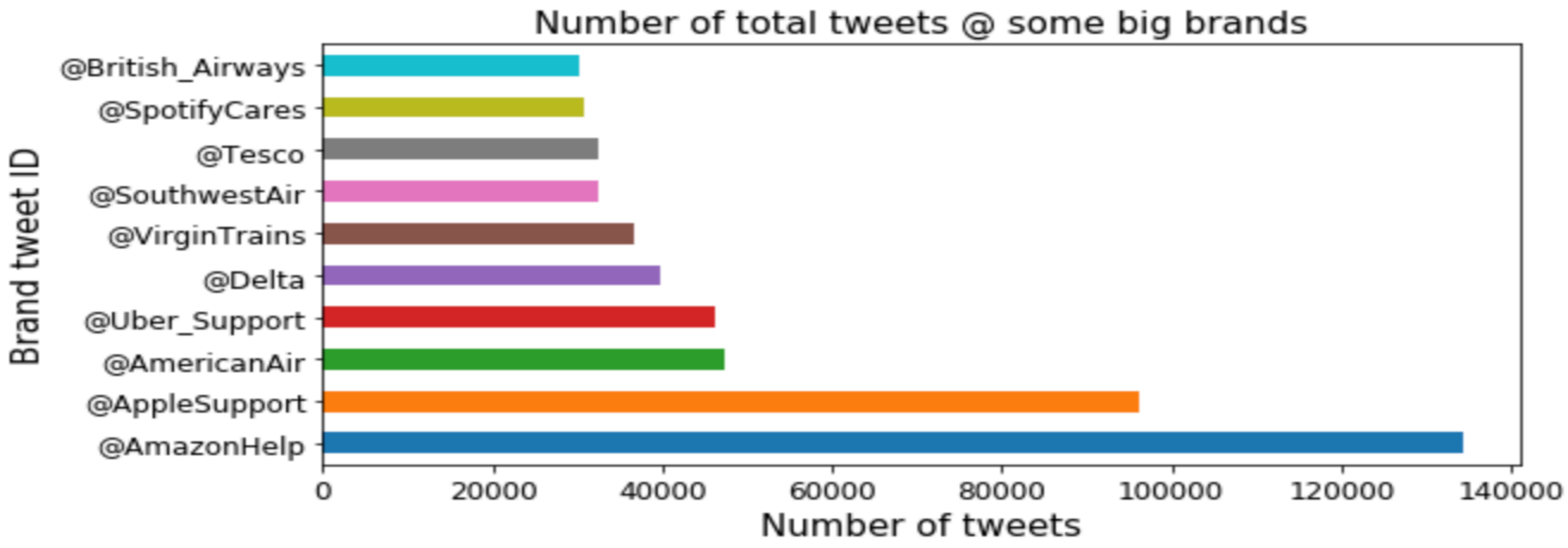
The purpose of the project is to understand what questions customers and ask and how to respond to them.

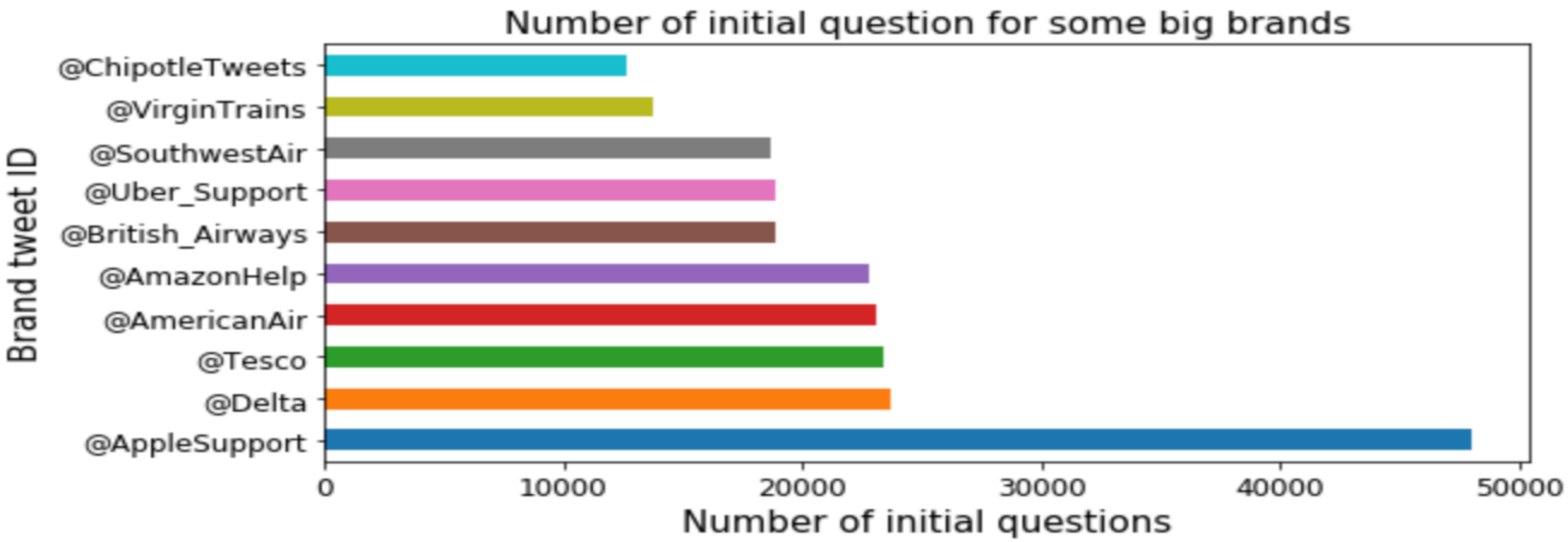
Data

- Kaggle dataset accessed from: https://www.kaggle.com/thoughtvector/customer-support-on-twitter

	tweet_id	author_id	inbound	created_at	text	response_tweet_id	in_response_to_tweet_id
0	1	sprintcare	False	Tue Oct 31 22:10:47 +0000 2017	@115712 I understand. I would like to assist y...	2	3.0
1	2	115712	True	Tue Oct 31 22:11:45 +0000 2017	@sprintcare and how do you propose we do that	NaN	1.0
2	3	115712	True	Tue Oct 31 22:08:27 +0000 2017	@sprintcare I have sent several private messag...	1	4.0
3	4	sprintcare	False	Tue Oct 31 21:54:49 +0000 2017	@115712 Please send us a Private Message so th...	3	5.0
4	5	115712	True	Tue Oct 31 21:49:35 +0000 2017	@sprintcare I did.	4	6.0

- 2811774 tweets from 4/1/2016 to 9/28/2016

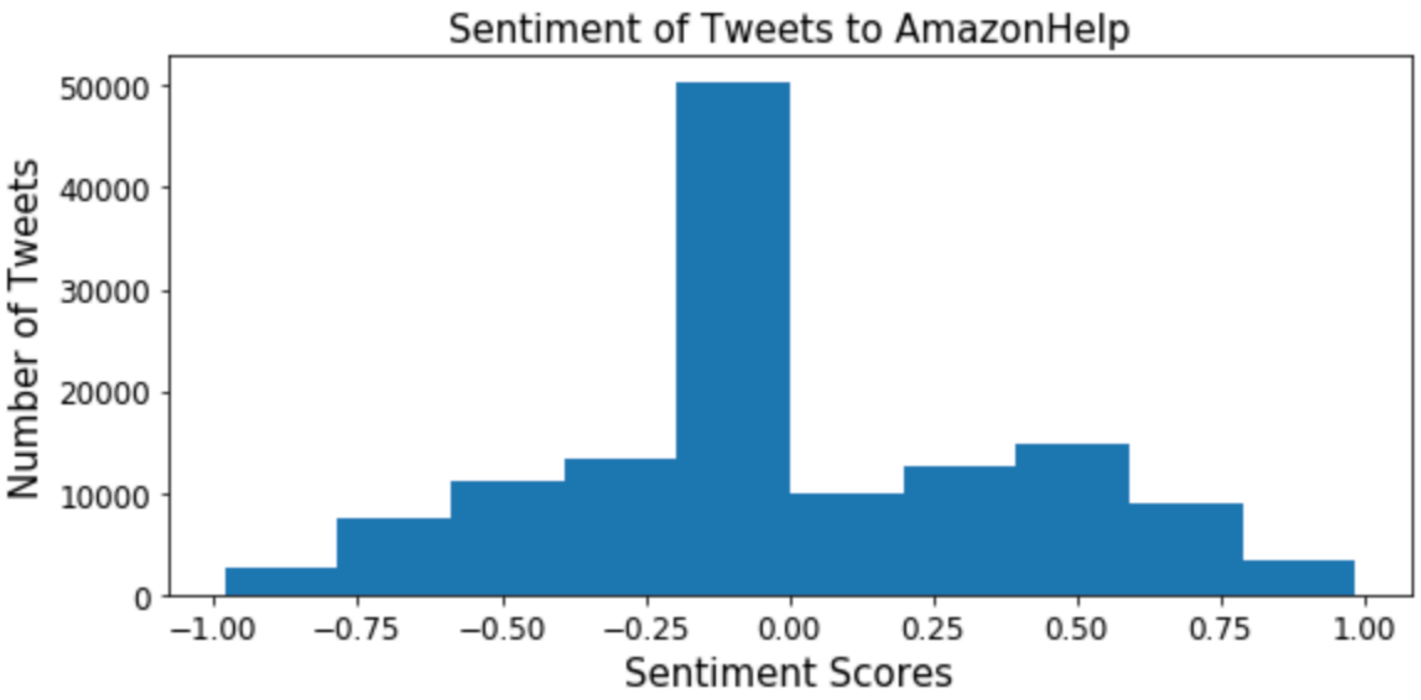


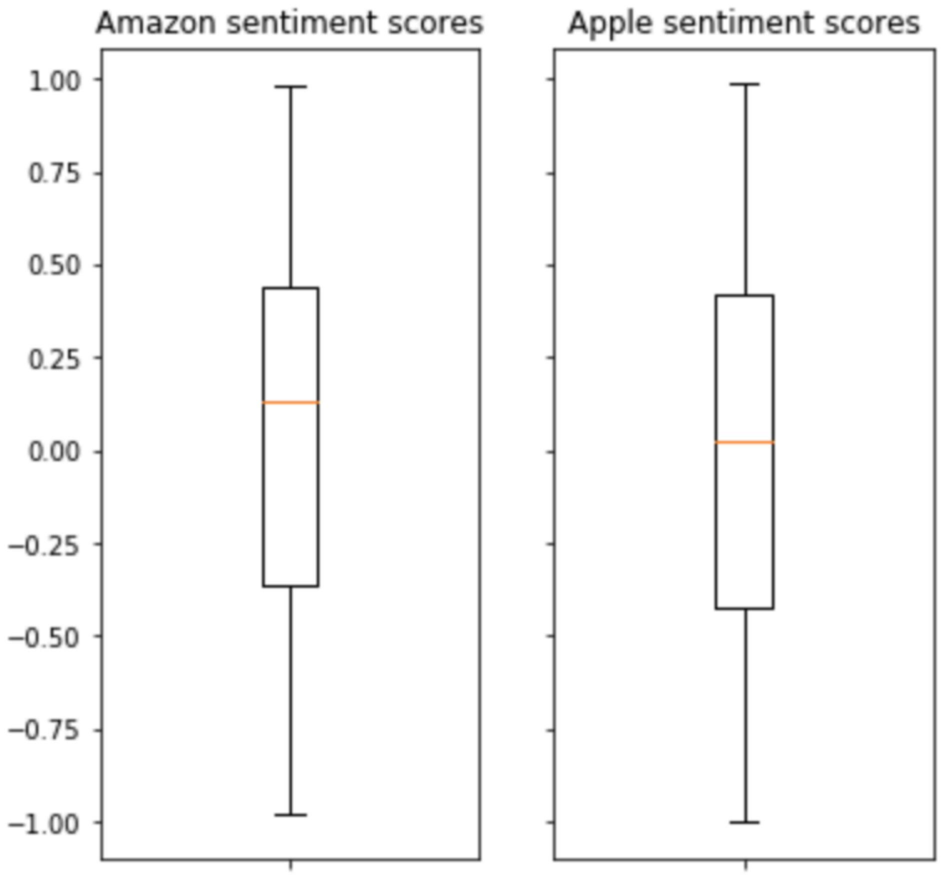


Only top 10 brands are included in the plot.

AppleSupport received the most number of initial questions /comments.

But, conversations are the most active at AmazonHelp.





Interpretation:
Calculating sentiment scores using vaderSentiment library. There are about the same number of positive and negative comments/questions @ AmazonHelp. And AmazonHelp in overall, has lightly more positive comments than AppleSupport.

Topics people asked

- Topic modeling methods: **LDA** and **NMF**
- Result: LDA indeed does a bad job on tweets (short texts)
- So, 10 topics in the questions interpreted from NMF top 20 words are:

1. placed order - shipping/delivery time
2. payment method (credit/gift card)
3. phone number (missing?)
4. wrong address /personal information changed
5. unknown (foreign language)
6. received product issue (items missing, damaged, returned, etc)
7. cancellation, kindle/membership
8. really disappointed of service (what amazon said, packages late, etc)
9. delay in order
10. customer support channel

Predicting a response to an initial question

- Modified existing sequence to sequence model
- Result:

```
Question

'Item has not been delivered but tracking says it was handed to me over an hour ago... 2nd time this has happened. So
rt it out https://t.co/42W82GcARk'

Predicted response from sequence to sequence model
```

Fitted on 10,000 tweets

```
'hi there , we are here to help . send us a note here ; and our team will follow up .'

Predicted response from sequence to sequence model
```

Fitted on 30,000 tweets

```
"hi there , i'm sorry to hear this . please dm us your email address and we'll be happy to help . ^ mm https://t.co/c
ooxeduuuc"

Actual response

"I'm so sorry you didn't receive your parcel! We'd like a chance to look into this with you here: https://t.co/JzP7hl
A23B ^SY"
```

Challenges

- Data reshaping: matching questions and responses.
- Data cleaning: filtering out numbers, links and most importantly foreign languages before topic modeling. I was able to remove most of foreign words, but there are still some remained.
  - Used enchant library to check language and TextBlob to translate (Not good enough).
  - Different cleaning steps for difference modeling purpose.
- Sequence to sequence model: hard to understand and took really long to run. I made the decision to move to AWS in the last day.