What financial topics do consumers discuss on social media?

After reading a few hundred messages, I noticed that a large percentage of the dataset provided no insights to answer this question. Many messages were either not from consumers or not related to financial topics. I performed a series of 'prunes' on the dataset that removed irrelevant messages. All code for this section can be found in pruning.ipynb.

Housekeeping Prunes:

1. Duplicate messages – *6.3% of original set*

Many messages had the same text as other messages in the dataset. All duplicate messages were removed in this prune.

2. Messages that could not be attributed to a bank - *9.1% of original set*

If no mentions, twitter handles, or hashtags for banks A-D were present in a message, it was removed in this prune.

Eliminating messages from non-consumers:

3. Bank responses to customers – *1.6% of original set*

I noticed that if 'Name_Resp' appeared at the end of a message, it was always a customer service representative responding to a customer. 'Name_Resp' wasn't mentioned in the metadata document, but I'm assuming this is a replacement for the twitter handle or name the representative is responding to. All messages including the string 'Name_Resp' were removed in this prune.

4. Interview Candidates, Employees, and HR - *.36% of original set*

Although interview candidates and employees are likely consumers of the banks they mention, the content of these messages did not provide information about consumer interaction with banks A-D. Posts about job openings also did not provide helpful information. Messages mentioning any of the words ['interview', 'interviews', 'got the job', 'hiring at', 'hired'] were removed in this prune.

5. News and Advertising – *45.1% of original set*

News and advertising posts were a huge percentage of the dataset. To filter these posts out, I fit a supervised classifier (logistic regression) to a collection of messages I labeled manually. The code I wrote to label is found in labeling.ipynb. My labeled set was 400 messages selected randomly from the already pruned dataset – bank responses, employees, sports and sponsored buildings, mission main street spam, and #getcollegeready posts were already removed. I manually categorized messages into two categories:

a. news – Definitely news or advertising *(125 of 400 labeled)*

b. not news – All other posts *(275 of 400 labeled)*

I fit a logistic regression classifier to these labeled messages, using L2 regularization. Although the messages are already tokenized, I added a few tokenizations to improve results, most notably grouping all bank mentions into one token for mentions, one for hashtags, and one for twitter handles. This prevented individual banks from becoming terms in my feature vector. I used 1 and 2 word terms in my model, removing terms if they did not occur in at least two messages.

accuracy=0.73500

	precision	recall	f1-score	support
N/A	0.73	0.88	0.80	207
neg	0.75	0.70	0.72	149
pos	0.70	0.16	0.26	44
avg / total	0.73	0.73	0.71	400

Precision, recall, f1-score for model 1

This model had a really high precision. Upon inspection of the classifier labeled messages, read about 200 labeled as 'news' and every message was news or advertising. However, a few news and advertising messages remained in the 'not news' labeled set. This is likely a result of being very strict with the 'news' label during hand labeling. The remaining news tweets will be removed in a later step.

Eliminating messages about non-financial topics:

5. Sports Arenas and Sponsored Buildings – *2.4% of original set*

Visitors to these venues may be customers, but they're talking about the game or event they're attending, not financial topics. I removed all messages mentioning any of the words ['stadium', 'playoffs', 'arena', 'preseason', 'center for the arts', 'BankA center', 'BankA building', 'game day'] in this prune.

6. 'Mission Main Street' – *2.5% of original set*

These messages are related to voting for small businesses in Chase's Mission Main Street Grants contest. They don't provide information about financial topics, so this prune removed all messages including the string 'mission main street'.

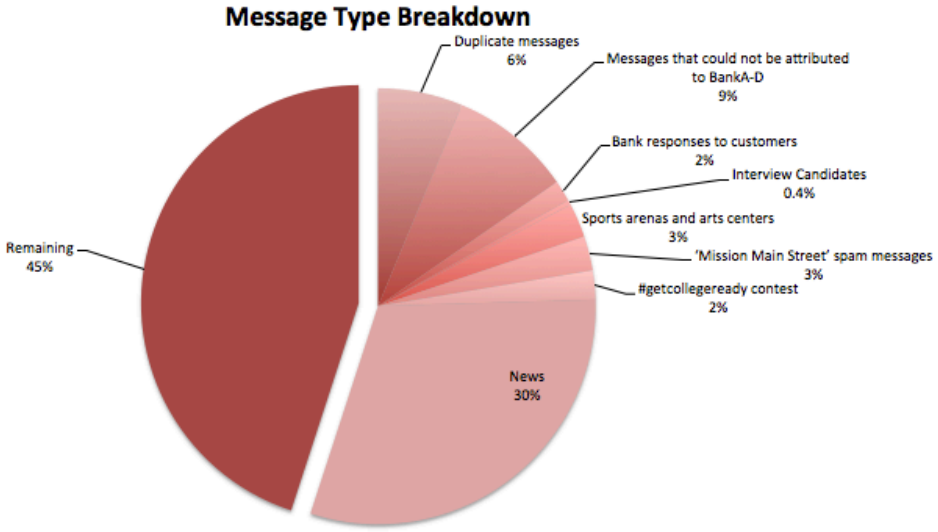
7. #getcollegeready – *2.2% of original set*

I removed these messages because they consisted of people participating in the contest and Wells Fargo responding to participants. It would be interesting to look more at these messages and the 'Mission Main Street' messages, but I think that analysis is out of the scope of this challenge. Because consumers are not discussing financial topics in these messages, I removed all messages with the string 'getcollegeready' in this prune.

The table below is a summary shows the type of message I removed, how many tweets I removed, and what percent of the initial set was removed.

Type of Message	**# of Tweets Removed**	**% of Original Set**
Duplicate messages	13846	6.3%
Messages that could not be attributed to BankA-D	19993	9.1%
Bank responses to customers	3448	1.6%
Interview Candidates	792	.36%
Sports arenas and arts centers	5307	2.4%
'Mission Main Street' spam messages	5606	2.5%
#getcollegeready contest	4757	2.2%
News and advertising	67226	45.1%
TOTALS	**120975**	**54.9%**

45.1% (99402 messages) remained after these removals.



Sentiment Classification:

After pruning, I hand labeled another 400 messages from the remaining 30% for sentiment. I manually categorized the messages into three categories:

a. Negative – These messages were from a customer expressing negative sentiment about an interaction with one of the four banks

b. Positive – These messages were from a customer expressing positive sentiment about a bank interaction

c. N/A – These messages provided no opinion on a financial topic. These messages are from users about financial topics, but they don't express sentiment.

I used the same type of logistic regression classifier as used in the news article classification. This model was great at distinguishing between N/A messages and sentiment messages, but not as good at distinguishing between positive and negative messages.

accuracy=0.73500

	precision	recall	f1-score	support
N/A	0.73	0.88	0.80	207
neg	0.75	0.70	0.72	149
pos	0.70	0.16	0.26	44
avg / total	0.73	0.73	0.71	400

Precision, recall, f1-score for model 2

I think this occurred because many of the messages were sarcastic, especially the tweets. A great example:

```
thanks to twit_hndl_BankB for refusing to exchange currency. excellent customer service.-
```

Considering that the second highest weighted positive term was 'thanks', it's understandable that the model had a difficult time recognizing the difference between positive and negative. More labels may have helped.

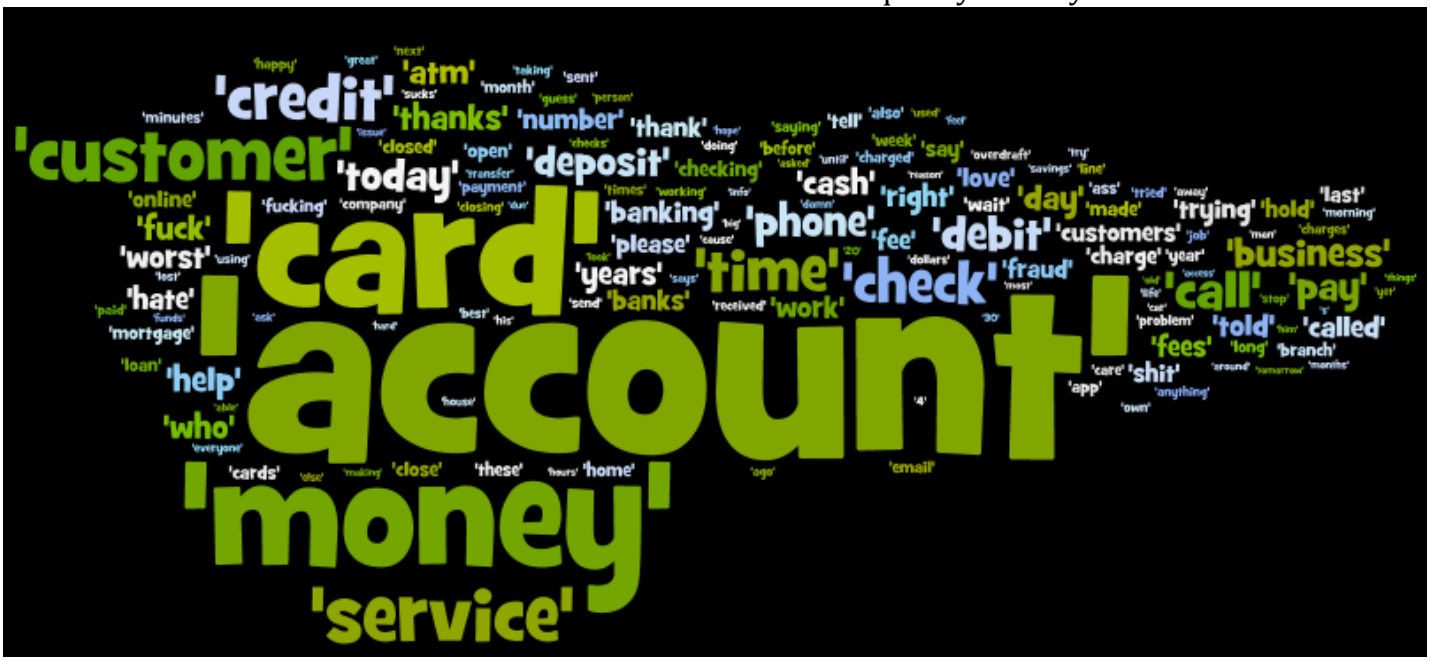
```
label distribution on messages: [('N/A', 60541), ('neg', 36772), ('pos', 2089)]
```

One important note is that the number of negative messages outweighs the number of positive messages by a factor of 17.6! Although the model isn't perfect, it's clear that consumers prefer to express negative feelings about their banking experience instead of positive feelings.

After this classification, the dataset was reduced to 17.6% of its original size (38861 messages). Code for labeling can be found in labeling.ipynb and code for classifying can be found in sent-prune.ipynb.

Topics:

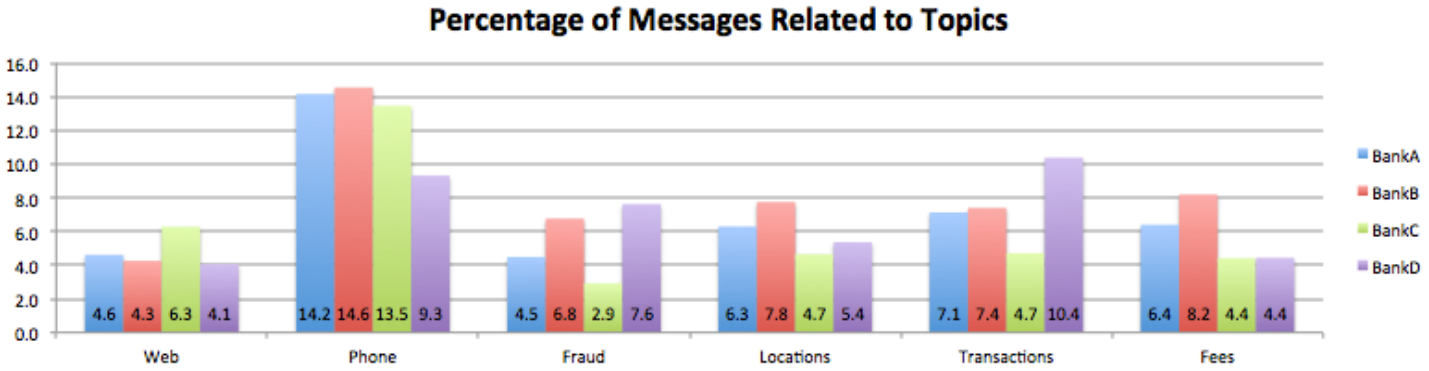
To determine topics mentioned, I counted the frequency of words in the remaining messages and used the significant words that occurred most frequently to make groupings for topic classification. This word cloud shows the list of words I looked at and their frequency visually:



The groupings I used were:

1: Web: ['online',' app ','website','login', 'logging in','mobile', 'password']
2: Phone: ['on hold', 'phone', 'mins', 'talk', 'call', 'automated', 'spoke to', 'speak', 'waiting', 'hung up']
3: Fraud: ['fraud', 'scam', 'hack', 'security', 'suspicious']
4: Physical Locations: ['branch', 'location', 'atm', 'teller', 'closing', 'town']
5: Transactions: ['deposit', 'withdrawl', 'payment', 'transaction', 'transferred']
6: Fees: ['fee', 'overdraft', 'charged']

I calculated the percentage of messages related to each bank that fell into these categories to make the table shown below. This table provides a lot of insight into the issues customers face when banking. Twice as many customers complained about their experience over the phone than any other category.



My next step would have been to take a second stab at sentiment classification with the smaller dataset. I actually did take a stab at it, but the numbers were pretty bad. Not only was accuracy super low, but I would have needed to label a lot more tweets to correctly label positives, which could lead to the model to be overfit.

accuracy=0.62121

	precision	recall	f1-score	support
neg	0.64	0.95	0.76	121
neutral	0.57	0.14	0.23	57
pos	0.00	0.00	0.00	20
avg / total	0.55	0.62	0.53	198

Precision, recall, f1-score for model 3

I would have loved to make a similar bar graph to the one above, but with each bar double tiered showing positive and negative sentiment as percentages of total messages with sentiment.

It's unfortunate that this challenge is so close to the end of the semester and during Thanksgiving. I had a blast analyzing the data and I wish that I had more time to submit a complete paper that met all requirements. I had a lot of ideas, but the implementation took longer than expected. My hope was to take a second stab at sentiment classifying within the topics to show which banks excelled or failed in each category. I also wanted to connect more information with my analysis – for example, it would have been great to talk about how the security breaches over the past year affected tweets for certain banks in the fraud category. I also counted links between banks (ie. BankA and BankB mentioned in the same message) and I wanted to explore whether customers dissatisfied with one bank were more likely to switch to another. Here are the numbers I got:

Bank Link	# of Messages	%XY
AB	1299	0.0447%
AC	217	0.0109%
AD	524	0.0235%
BC	424	0.0224%
BD	807	0.0380%
CD	281	0.0244%

$$\%XY = \frac{\#\ of\ msgs\ \ with\ X\ and\ Y}{(\#of\ msgs\ with\ X\ only) + (\#\ of\ msgs\ with\ Y\ only)}$$

Thanks for a cool challenge!