Project 1 Presentation

Ryan Downing, Stefan Obradovic, Sahil Goel, Candace Austin, Rohan Samy, Kafana Ouattara

Dataset Description

We collected 313 tweets from individuals on Twitter regarding the technology sector

- Twitter posts made between March 01, 2020 and March 12, 2022
- At most 10 tweets per analyst
- Tweets specifically mentioning the following tickers: AAPL, MSFT, GOOG, AMZN, TSLA, FB, NVDA, AVGO, CSCO, ADBE, CRM, ORCL, ITC, AMD, NFLX, TXN, INTU, PYPL, NOW, IBM, QQQ, TQQQ, ARKK

We then hand-labelled each tweet as having either positive sentiment (1) or non-positive sentiment (0) in an effort to analyze public opinion about the tech sector online during this period.

Collecting the Dataset

Used Python's tweepy Twitter API

- Collected list of twitter usernames from active users in the financial analyst community
- Obtained all tweets for each user in the set time range
- Filtered tweets which did not contain any cashtags (i.e. $AAPL, $MSFT)

```python
def get_users_tweets_between_dates(self, user, start_date, end_date, **kwargs):
        start_date_str = start_date.replace(tzinfo=pytz.UTC).isoformat()
        end_date_str = end_date.replace(tzinfo=pytz.UTC).isoformat()
        tweets = []
        while True:
            new_tweets = self.get_users_tweets(
              user, max_results=100, end_time=end_date_str,
              tweet_fields=["id", "text", "created_at", "author_id"], **kwargs
            )
            if new_tweets is None:
                return tweets

            new_tweets = list(reversed(new_tweets))
            if len(new_tweets) < 100:
                return tweets + new_tweets

            oldest_tweet_date = new_tweets[0].created_at.replace(tzinfo=None)
            if oldest_tweet_date < start_date:
                for i, tweet in enumerate(new_tweets, 1):
                    date = tweet.created_at.replace(tzinfo=None)
                    if date >= start_date:
                        return tweets + new_tweets[i:]
                return tweets

            end_date_str = oldest_tweet_date.replace(tzinfo=pytz.UTC).isoformat()
            tweets.extend(new_tweets)
```

Author	author_id	created_at	id	Text	Date	Stock(s) Mentioned	Tweet Link	Misc.
DanielTNiles	1948086848	2020-05-04 03:49:04+00:00	1257155255852085249	On @CNBC talking about managing positions dail...	2020-05-04 03:49:04+00:00	AMZN	https://twitter.com/twitter/statuses/125715525...	
DanielTNiles	1948086848	2020-07-17 19:21:26+00:00	1284206597179228160	On @YahooFinance w/ @zGuz about rotating into ...	2020-07-17 19:21:26+00:00	TSLA	https://twitter.com/twitter/statuses/128420659...	
DanielTNiles	1948086848	2021-05-12 15:24:59+00:00	1392501064193122304	As I said on @CNBC, I still like $VIAC. Q1 res...	2021-05-12 15:24:59+00:00	NFLX	https://twitter.com/twitter/statuses/139250106...	
DanielTNiles	1948086848	2020-07-30 18:32:12+00:00	1288905248241676288	Covering $AMZN short & getting long again....	2020-07-30 18:32:12+00:00	MSFT, AMZN	https://twitter.com/twitter/statuses/128890524...	
DanielTNiles	1948086848	2020-10-20 20:25:34+00:00	1318649583405309953	The power of social media/internet meeting tal...	2020-10-20 20:25:34+00:00	FB	https://twitter.com/twitter/statuses/131864958...	
...	...	...	...	...	...	...	...	...
The_RockTrading	21764428	2022-03-05 16:40:00+00:00	1500149136645050377	$FB Closing Friday below critical $200 support...	2022-03-05 16:40:00+00:00	FB	https://twitter.com/twitter/statuses/150014913...	
The_RockTrading	21764428	2022-02-16 13:13:23+00:00	1493936546591870986	$FB 👀	2022-02-16 13:13:23+00:00	FB	https://twitter.com/twitter/statuses/149393654...	
The_RockTrading	21764428	2022-03-05 16:20:00+00:00	1500144103488737288	$TSLA Doesn't look terrible like the rest of t...	2022-03-05 16:20:00+00:00	TSLA	https://twitter.com/twitter/statuses/150014410...	

Punctuation, Numbers, and Syntax

Cleaning text before applying analysis

- Removed all characters which are not alphabetical or whitespace (including emojis and cashtags)
- Use NLTK's porter stemmer to get the roots of all words
- Split on whitespace to get tokens, remove stopwords, and recreate sentence

```
stemmer = PorterStemmer()
sw = set(stopwords.words("english"))

def stem_and_remove_sw(tokens):
  return [stemmer.stem(token) for token in tokens if stemmer.stem(token) not in sw]

data["cleaned_text"] = data["Text"].str.replace("[^A-Za-z\s]", "", regex=True).str.lower()

data['tokens'] = data["cleaned_text"].str.split(',')
data['tokens'] = data['tokens'].apply(stem_and_remove_sw)
data["tokenized_sentence"] = data["tokens"].apply(lambda tokens: ' '.join(tokens))
```

TF-IDF

Constructing the Term Frequency - Inverse Document Frequency (TF-IDF) matrix:

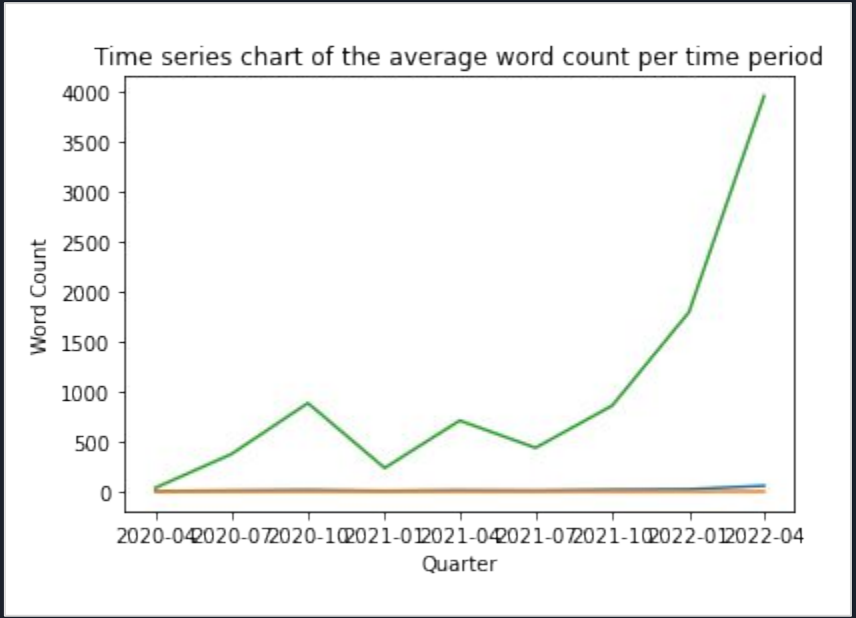
```
# getting the TF-IDF matrix
tfidf_vectorizer = TfidfVectorizer(use_idf=True)
X = tfidf_vectorizer.fit_transform(data["tokenized_sentence"])
tweets_df = pd.DataFrame(columns = tfidf_vectorizer.get_feature_names(), data = X.toarray())
tweets_df
```

- Using SciKit Learn library and applying it to the tokenized tweets, we created a 318 by 2624 matrix with TF-IDF weights
- Each row corresponds to a tweet (document) and each column represents a unique word in the corpus

Analysis:

- Highest TF measure: 6 (from the term "early" in document 11)
- Highest IDF measure: 6.072 (from the term "def")
- The highest TF-IDF measure: 1 (from the term "fb" in document 314)

Time Series Chart

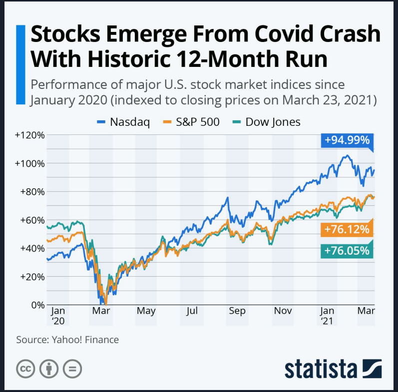


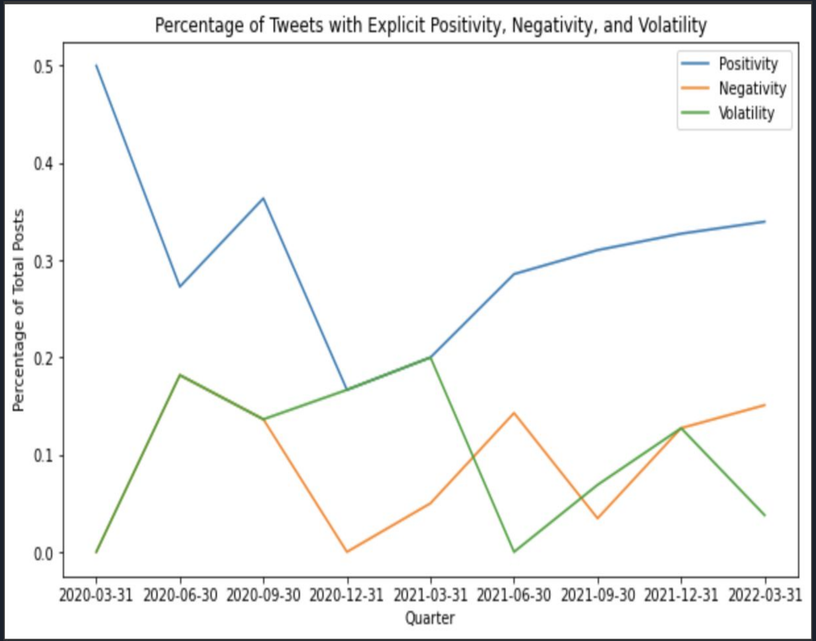
Regex Count

Terms were chosen after manual labelling and searching for stock-related sentiment words. Calculated percentage of tweets by quarter that mention at least one of the following terms

- Positivity Words: `good|well|outperform|up|rise|increase|bull`
- Negativity Words: `bad|underperform|down|drop|fall|decline|bear`
- Volatility Words: `large|uncertain|volatility|huge|spike|unsure|risk`

- With the exception of Q1 2020 and Q3 2021, positivity and negativity follow same trend (likely because controversy/news stirs both sides).
- Overall, more positive sentiment on twitter (likely corresponds with general market recovery)
- Higher mention of market volatility in earlier months, sooner into the market recovery from covid-19.





Naive Bayes Classifier- TF-IDF Matrix

Ran two Naive Bayes Classifiers:

- One using the Term Frequency-Inverse Document Frequency Matrix
- Another one using the Count Document Matrix

Word Indicative of 1 (Positive):

- money
- splits
- hip
- new
- values
- climbs
- good
- biggest
- expand
- outperform
- postearnings

Word Indicative of 0 (Non-Positive):

- shortage
- short
- outlier
- micro
- hitting
- turning
- ridiculous
- otherwise
- not
- rejection
- discount

```
print(Positive_Words)
print(NonPositive_Words)
```

```
     Positive Words
0            abdiel
1             above
2          absolute
3        accelerate
4         according
..              ...
959            youl
960            your
961             yoy
962             yrs
963              yy

[964 rows x 1 columns]
      Non-Positive Words
0                   aaii
1                   aapl
2                  about
3                  abuse
4               accepted
...                  ...
1655                  zm
1656                zone
1657                zoom
1658                  zs
1659                zuck

[1660 rows x 1 columns]
```

```
import random
print((random.sample(pos_words, 10)))
print((random.sample(nonpos_words, 10)))
```

```
['cos', 'outperform', 'jd', 'touched', 'postearnings', 'among', 'stays', 'end', 'xom', 'staying']
['not', 'first', 'rejection', 'discount', 'who', 'twlo', 'gld', 'techs', 'webinar', 'terrific']
```

Accuracy Score: 0.725

Confusion Matrix [80 Observations]

11 - Correctly Labeled Positive
5- False Positive
17- False Non-Positives
47- Correctly Labeled False

Naive Bayes Classifier- Count Document Matrix

Ran two Naive Bayes Classifiers:

- One using the Term Frequency-Inverse Document Frequency Matrix
- Another one using the Count Document Matrix

Word Indicative of 1 (Positive):

- great
- economy
- valuation
- held
- technology
- central
- expecting
- positive
- rewarded
- privacy
- market

Word Indicative of 0 (Non-Positive):

- accountability
- shortages
- term
- cycle
- opportunities
- stops
- delivery
- premium
- panicked
- blamed
- liability

```
    Positive Words
0           abdiel
1         absolute
2       accelerate
3        according
4       accounting
..             ...
935          youll
936            yoy
937             yr
938            yrs
939             yy

[940 rows x 1 columns]
     Non-Positive Words
0                  aaii
1                  aapl
2                 abuse
3              accepted
4           accordingly
...                 ...
1555                 zm
1556               zone
1557               zoom
1558                 zs
1559               zuck

[1560 rows x 1 columns]
```

Accuracy Score: 0.7125

Confusion Matrix [80 Observations]

16 - Correctly Labeled Positive

11 - False Positive

12 - False Non-Positives

41 - Correctly Labeled False

```
import random
print((random.sample(posi_words, 10)))
print((random.sample(non_posi_words, 10)))

['great', 'economy', 'delivered', 'mlb', 'steve', 'os', 'decadewe', 'values', 'innovation', 'opinion']
['accountability', 'defacto', 'pairings', 'clear', 'street', 'using', 'found', 'establishment', 'quality', 'would']
```

Paper Summary - Background

- Authors wanted to compare the effect of COVID-19 on stock market volatility with that of the Spanish Flu of 1918, the Influenza Pandemic of 1957–1958, and the Influenza Pandemic of 1968

- They noticed that from February 24, 2020 through April 2020, COVID-19 news drove volatility (next-day newspaper accounts attribute 23 of 24 of most volatile days in markets to news about COVID-19 developments and policy responses to the pandemic)

- Through textual analysis of news articles, authors try to explain the rationale for what makes this pandemic's effect on the stock market so much greater than the previous pandemics

The Review of Asset Pricing Studies

The Unprecedented Stock Market Reaction to COVID-19

Scott R. Baker
Kellogg School of Management, Northwestern University

Nicholas Bloom
Stanford University

Steven J. Davis
Chicago School of Business, University of Chicago

Kyle Kost
University of Chicago

Marco Sammon
Kellogg School of Management, Northwestern University

Tasaneeya Viratyosin
Wharton School of Business, University of Pennsylvania

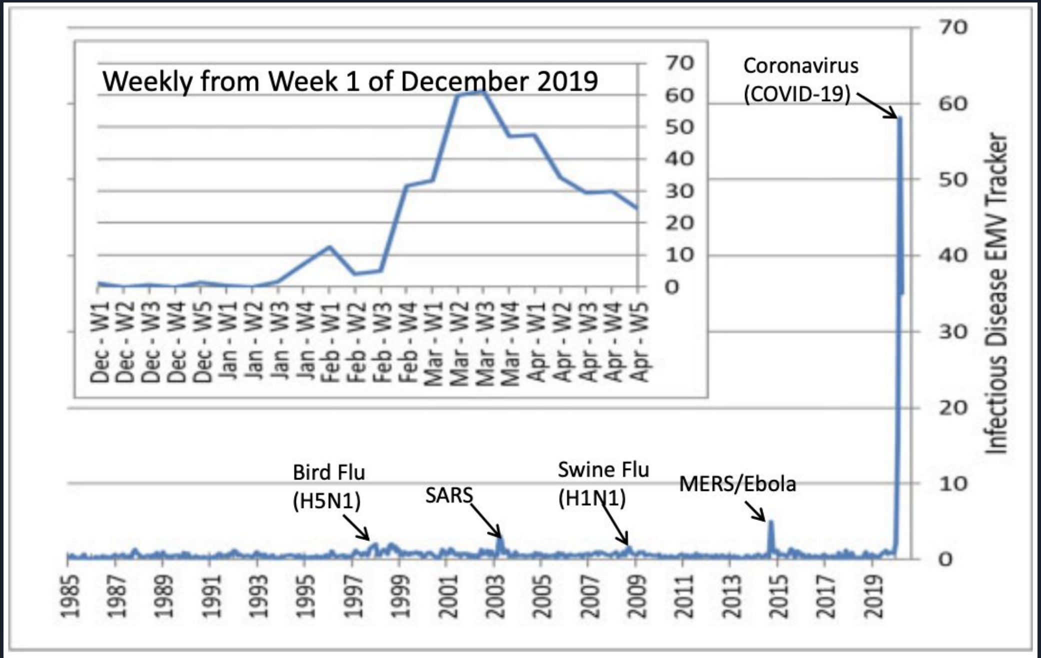
Paper Summary - Methodology

Quantifying the Contribution of COVID-19:

- Calculate fraction of articles that mention terms related to market or volatility
- Scale to mean value of VIX
- Get subset of articles that mention a set of infectious diseases

Conclusions:

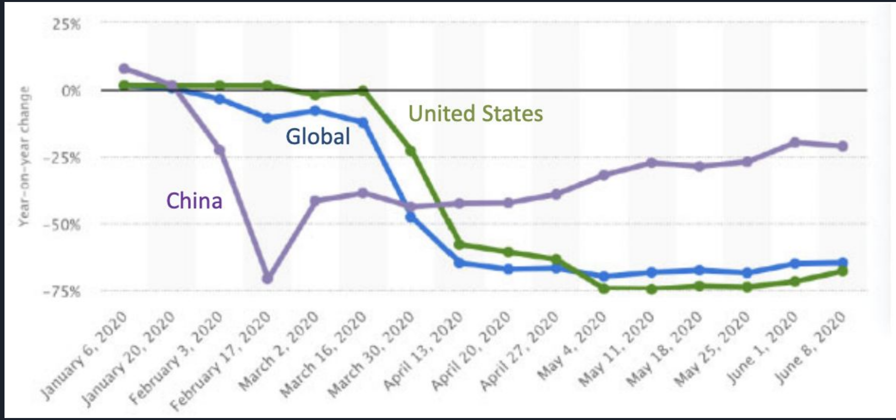
- Before COVID-19, no infectious disease outbreak made a sizable contribution to U.S. stock market volatility
- COVID-19 pandemic drove the tremendous recent surge in stock market volatility
- COVID-19 volatility surge began in the fourth week of January, intensified from the fourth week of February, and began tapering in the fourth week of March



Paper Summary - Possible Explanations

Unlikely to explain large effects on stock market:

- Ease with which the virus spreads, and the non negligible mortality rate
- Information about pandemics is richer and diffuses much more rapidly now than a century earlier
- Cross-border flows of goods in the modern economy



Likely to explain large effects on stock market:

- high-volume international travel and the predominant role of the service sector
- Nonpharmaceutical policy interventions (NPIs)
  - Travel Restrictions
  - Stay-at-home orders, school & restaurant closures
  - Covid relief money
- Voluntary declines in social activity

Paper Summary - Conclusions

- Stock market volatility largely caused by
    - Mandatory business closures
    - Restrictions on commercial activity
    - Voluntary social distancing
    - Service-oriented economy

- Unlike before, government restrictions on commercial activity in response to COVID-19 were more stringent, broader in scope, more widespread, and lengthier in duration

- There is a compelling need to address the health crisis created by COVID-19, while shifting to less sweeping containment policies that do not strangle the economy

Thank You!!