

# Data607 Final Project

2019 Trump Tweets

# Big Bang Team



- Sie Siong Wong
- Anil Akyildirim
- Joe Rovalino



Can we leverage President Donald Trump's Tweets to predict the stock market?

# Agenda

- Organizational
- Lifecycle
  - Team formation, Decision on topic of interest
  - Data Acquisition, Data cleanup, Topic model, sentiment creation
  - Analysis of sentiment of tweets vs market data
- Conclusions
- Wrap up

# Organize

- Slack Private Channel
  - Collaboration Tool
  - Asynchronous - group threads and breakout sessions
- Skype
  - Voice and video
- Github
  - Version Control

# Lifecycle

- Formed Team
- Explored question(s) of interest/decision on topic --> Trump Tweets
- Researched Topic model approaches
- Tested Connectivity to Twitter API and package twitterR
- Downloaded Tweet data and S&P market data
- Performed data analysis and data cleanup

# Lifecycle (cont.) -

- Consulted with Professor on approach
- Aggregated tweets by date to create sentiment/measure
- Compared Tweet sentiment and S&P performance
  - Tweets tokenized/frequency of words
  - Cleanup corpus
- Prepared Visualizations to assist with conclusion analysis

# Lifecycle (cont.) -

- Compared Tweet sentiment and S&P performance (cont.)
  - Tweets tokenized/frequency of words
  - Cleanup corpus
  - Create and filter Document Term Matrix (DTM)
  - DTM exploration
- Prepared Visualizations to words



# Lifecycle (cont.) -

- Visualization of stock data trends
- Applied LDA Topic Model
  - Identify the topics of interest
  - Create LDA model using parameters Gibbs method for 30 topics
  - Per Document classification
- Sentiment Analysis
- Sentiment score vs stock price change

# Data Collection

## Stock Market Data

- Yahoo Finance Stock Data
- Export csv and load to R



## Tweets Data

- Twitter Developer Account
- API Key and Token
- 3200 tweets limit
- Trump Twitter Archive
- Export csv and load to R

```
# Authorization keys.
app_name <- "JAS"
consumer_key <- "sPwbbZCtf8nfs1xhYtZqI8MHJ"
consumer_secret <- "KfcOxgElcQ70f13QMy8LkuDAN18dunXT147H0A8a80Lzpr3Vd3"
access_token <- "600477513-rdd3Fcyuq1sfnh5560egRQxx0TIDqfrLzyZo4Vik"
access_secret <- "SdDFCJU0oqAw671VXeLa0781TdUvde8Sh2gyQW64P5Zh"

# Extract some tweets from Twitter.
setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)

## [1] "Using direct authentication"

tweets <- userTimeline("realDonaldTrump", n=5)
tweets

## [[1]]
## [1] "realDonaldTrump: Without the horror show that is the Radical Left, Do Nothing Democrats, the Stock Markets and Economy would be even... https://t.co/em6px5e20D"
```

```
# President Trump tweets from 01/01/2018 to 11/21/2019.
tweets_raw <- read.csv("https://raw.githubusercontent.com/Siesiongiang/Twitter/dev/trumptweets.csv")

# SBP stock price data from year 01/04/2016 to 11/21/2019.
stocks_raw <- read.csv("https://raw.githubusercontent.com/Siesiongiang/Twitter/dev/sandp.csv")

head(tweets_raw)

##           source
## 1 Twitter for iPhone
## 2 Twitter for iPhone
## 3 Twitter for iPhone
## 4 Twitter for iPhone
## 5 Twitter for iPhone
## 6 Twitter for iPhone
##
## text
## 1
## https://t.co/osezCwP01                                     Poll: Trump leads top 2020 Democrats in
## 2 wisconsin https://t.co/872751820N.
## 3 RT @realDonaldTrump: Impeachment Witch Hunt is now OVER! Ambassador Sondland asks U.S. President (me): "W
## 4 RT @realDonaldTrump: ... "I WANT NOTHING! I WANT NOTHING! I WANT NO QUID PRO QUO! TELL PRESIDENT ZELENSK
## 5 "All four of Gordon Sondland's lawyers are Democrat Donors." @TuckerCarlson Despite
## 6 Watch @TuckerCarlson @seanhannity
## created_at retweet_count favorite_count is_retweet id_str
## 1 11/21/2019 2:47 24221 62863 false 1.1973464e+18
## 2 11/21/2019 1:22 14184 52661 false 1.1973244e+18
## 3 11/21/2019 1:16 23980 0 true 1.1973224e+18
## 4 11/21/2019 1:16 18754 0 true 1.1973224e+18
## 5 11/21/2019 1:11 16331 68155 false 1.1973224e+18
## 6 11/21/2019 1:03 9837 37964 false 1.1973204e+18
```

# Data Cleaning

## Stock Market Data

- Update the Date format.
- Select Date Range.
- Create Price Change.

```
# Update Date column into date format.
stocks_raw$Date <- as.Date(stocks_raw$Date)

# Select data from 01/01/2018 to 11/20/2019 and calculate price change percentage between close and open price.
stocks.df <- stocks_raw %>%
  filter(between(Date, as.Date("2018-01-01"), as.Date("2019-11-20"))) %>%
  mutate(Pct_Change=(Close-Open)/Open*100)

head(stocks.df)
```

## Tweets Data

- Select Required Columns.
- Separate Date and Hour.
- Remove meaningless characters.
- Remove hour from Date.
- Remove tweets less than 20 characters.
- Add id to each tweet.

```
# Drop source column.
tweets_slc <- tweets_slc %>% select(text, created_at)

# Separate column "created_at" into "date" and "hour".
tweets_slc <- separate(data = tweets_slc, col = created_at, into = c('date', 'hour'), sep = ' ') %>% select(text, date, hour)
```

```
# Remove minutes in hour column.
tweets_slc$hour <- gsub(":\\:\\w*", "", tweets_slc$hour)

# Remove meaningless characters and symbols.
tweets_slc$text <- gsub("&"," ", tweets_slc$text)
tweets_slc$text <- gsub("(RT)(?:\\b\\w*\\b\\w+)", "", tweets_slc$text)
tweets_slc$text <- gsub("^RT", "", tweets_slc$text)
tweets_slc$text <- gsub("@\\w+", "", tweets_slc$text)
tweets_slc$text <- gsub("[[:punct:]]", "", tweets_slc$text)
tweets_slc$text <- gsub("[[:digit:]]+\\s", "", tweets_slc$text)
tweets_slc$text <- gsub("http\\w+", "", tweets_slc$text)
tweets_slc$text <- gsub("[ \\t]{2,}", " ", tweets_slc$text)

# Remove all non-ASCII characters
tweets_slc$text <- iconv(tweets_slc$text, "UTF-8", "ASCII", sub="")

# Delete empty text column.
tweets_slc <- tweets_slc %>% na_if("") %>% na_if(" ") %>% na.omit()

# Tweets that contained less than 20 characters were treated as noise.
tweets_slc <- tweets_slc %>% filter(nchar(text)>20)

# Add id column to consider each text row as a document.
tweets_slc$doc_id <- seq.int(nrow(tweets_slc))
```

# Create and Clean Corpus

- Remove Stop Words (en and smart)
- Remove Numbers
- Remove White Space
- Remove Sparse Terms

```
# Select text and id column.
tweetscorpus.df <- tweets_slc %>% select(doc_id, text)

# Create a corpus for document term matrix.
tweetscorpus <- VCorpus(DataframeSource(tweetscorpus.df))

# Remove all punctuation from the corpus.
tweetscorpus <- tm_map(tweetscorpus, removePunctuation)

# Remove all English stopwords from the corpus.
tweetscorpus <- tm_map(tweetscorpus, removeWords, stopwords("en"))
tweetscorpus <- tm_map(tweetscorpus, removeWords, stopwords("SMART"))

# Remove all number from the corpus.
tweetscorpus <- tm_map(tweetscorpus, removeNumbers)

# Strip extra white spaces in the corpus.
tweetscorpus <- tm_map(tweetscorpus, stripWhitespace)

# Stem words in the corpus.
tweetscorpus <- tm_map(tweetscorpus, stemDocument)

# Build a document term matrix.
tweetsdtm <- DocumentTermMatrix(tweetscorpus)

# Remove sparse terms which don't appear very often. Limit the document term matrix to contain terms appearing
in at least 2% of documents.
tweetsdtm <- removeSparseTerms(tweetsdtm, 0.98)

# Find the sum of words in each document and remove all docs without words.
rowTotals <- apply(tweetsdtm, 1, sum)
tweetsdtm.new <- tweetsdtm[rowTotals > 0, ]

# Put the document in the format lda package required
tweetsdtm.matrix <- as.matrix(tweetsdtm.new)

head(tweetsdtm.matrix, n=5)
```

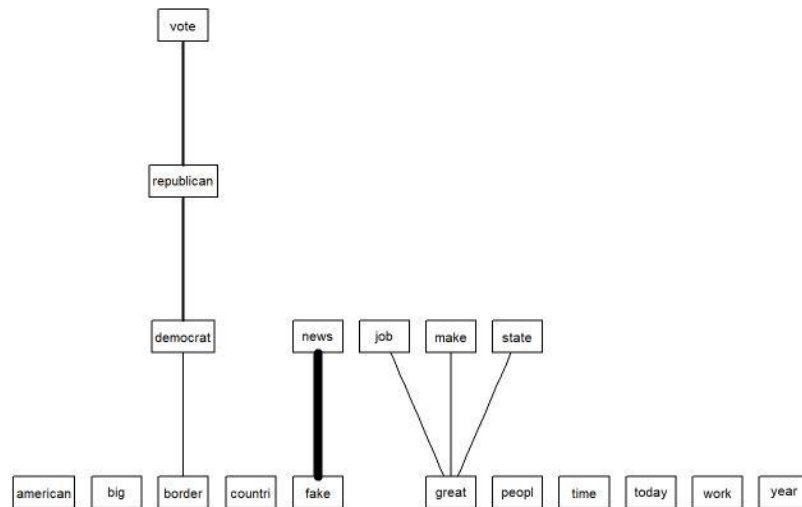


# Association of Words

```
## $china
## billion    deal    trade    dollar    continu    usa    good    meet    make
##    0.19    0.19    0.19    0.16    0.13    0.11    0.08    0.08    0.07
##    year    unit    start
##    0.07    0.06    0.05
```

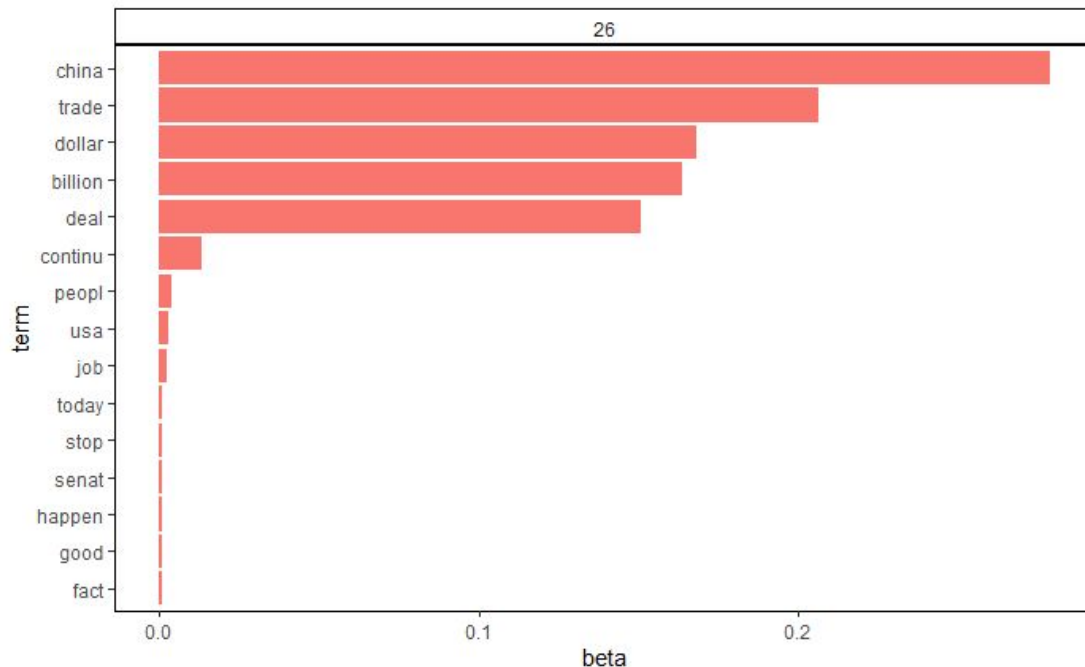
```
## $trade
##    deal billion    china    countri    dollar    year    unit    talk    good
##    0.25    0.20    0.19    0.13    0.13    0.12    0.10    0.08    0.06
##    usa    long    meet
##    0.06    0.05    0.05
```

```
## $job
##    great    militari    economi    record    number    tax
##    0.13    0.09    0.07    0.07    0.06    0.06
```



# LDA Topics Modeling - Gibbs Method

Topic 23	Topic 24	Topic 25	Topic 26
"border"	"vote"	"countri"	"china"
"wall"	"republican"	"histori"	"trade"
"secur"	"parti"	"world"	"dollar"
"vote"	"big"	"usa"	"billion"
"schiff"	"elect"	"trade"	"deal"
"hard"	"countri"	"make"	"continu"
"world"	"crime"	"democrat"	"peopl"
"deal"	"impeach"	"includ"	"usa"
"high"	"bad"	"republican"	"job"
"make"	"import"	"end"	"fact"
"nation"	"media"	"back"	"good"
"republican"	"russia"	"day"	"happen"
"today"	"trade"	"peopl"	"senat"
"administr"	"end"	"thing"	"stop"
"america"	"stori"	"long"	"today"
"american"	"talk"	"total"	"administr"
"back"	"back"	"news"	"america"
"bad"	"democrat"	"parti"	"american"
"big"	"happen"	"state"	"back"
"billion"	"hous"	"american"	"bad"
"call"	"illeg"	"border"	"big"
"campaign"	"meet"	"congratul"	"border"
"china"	"administr"	"hous"	"call"
"collus"	"america"	"import"	"campaign"
"congratul"	"american"	"meet"	"collus"
"congress"	"billion"	"report"	"congratul"
"continu"	"border"	"administr"	"congress"
"corrupt"	"call"	"america"	"corrupt"
"countri"	"campaign"	"bad"	"countri"
"crime"	"china"	"big"	"crime"
Topic 29	Topic 30		





# Per Document Classification

**text**  
<chr>

poll leads top democrats in wisconsin  
impeachment witch hunt is now over ambassador sondland  
i want nothing i want nothing i want no quid pro quo tell  
all four of gordon sondland's lawyers are democrat donors  
if this were a prizefight they'd stop it  
today i opened a major apple manufacturing plant in texas  
with every question that asks democrats sham impeachment  
obama gave ukraine blankets gave ukraine missiles  
i want nothing i want no quid pro quo i want zelensky to  
wow mr sondland let's be clear no one on this planet not c

1-10 of 9,157 rows | 1-1 of 4 columns



**text**  
<chr>

water levels could begin to rise well in advance of the arrival  
here's the latest dorian storm surge forecast from this morning  
the has issued a high risk area for flash flooding over eastern  
us winning trade war with china in dollars cnbc  
to declassify is so important because if this were a democrat  
since my election many trillions of dollars of worth has been  
am edt tropical cyclone update on hurricane dorian eye of do  
the free red cross emergency app puts realtime weather alert  
the euro is dropping against the dollar like crazy giving them  
the wall is going up very fast despite total obstruction by dem

41-50 of 253 rows | 3-3 of 5 columns

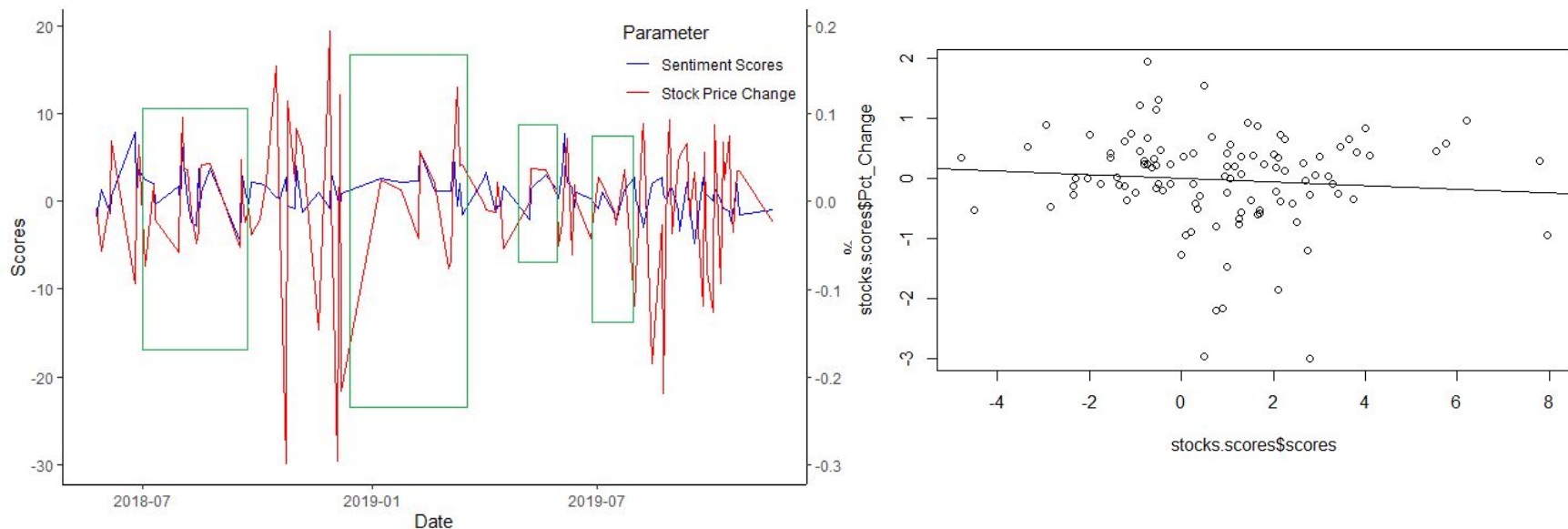


# Sentiment Analysis

anger	anticipation	disgust	fear	joy	sadness	surprise	trust	negative	positive
1	1	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0
2	1	0	3	1	2	1	4	4	5
1	0	1	0	0	0	0	2	2	0
0	1	0	0	1	0	1	1	1	2
1	0	1	0	0	1	0	0	1	0

topics.tweetsLDA	doc_id	text	Date	hour	tweets.score
26	61	i am struck by schiffs attempt to characterize s conversatio...	11/19/2019	18	-1.000000e+00
26	62	we all have the transcript of the call schiff is asking vindma...	11/19/2019	18	0.000000e+00
26	512	there is serious work to get done on behalf of this countrya...	11/2/2019	21	0.000000e+00
26	719	general michael flynns attorney is demanding that charges ...	10/26/2019	11	-1.500000e+00
26	786	democrats are trying to deny republican members of Congr...	10/23/2019	17	-1.550000e+00
26	865	doral in miami would have been the best place to hold the ...	10/21/2019	13	2.100000e+00
26	933	such a disgrace that the do nothing democrats are doing ju...	10/19/2019	15	-1.000000e+00

# Sentiment Scores vs Stock Price Change



Coefficients:





	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.002823	0.083193	-0.034	0.973
scores	-0.029931	0.034463	-0.869	0.387

Residual standard error: 0.8041 on 105 degrees of freedom

Multiple R-squared: 0.007133, Adjusted R-squared: -0.002323

# Conclusion

- Surprise! Slight negative linear correlation b/n the sentiment scores and percent change on stock market.

Date <date>	Pct_Change <dbl>	scores <dbl>
2018-05-25	-0.08334630	-1.75
2018-05-29	-0.56374785	1.30
2018-06-05	0.01237377 	-1.40 
2018-06-06	0.69372916	0.65
2018-06-25	-0.94314398 	7.95 
2018-06-26	0.03452978	0.95

- Sentiment Scores fluctuates more than the Percentage Change in Stock Market.
- Top 5 words - "china", "trade", "dollar", "billion" and "deal" - most common within the topic with most impact to stock market price change
- Explicit "trade" word in the tweet, the most common word/phrase are: "such as deal", "billion", "china", "countri", "dollar", "year", "unit", "talk", "good", "usa", "long" and "meet"

# Wrap Up

- Q and A
- Thank You!