Predicting S&P500 Volatility using Natural Language Processing and Statistical Learning Techniques

By Andrew Orf

Dr. Christopher Wikle, Project Supervisor

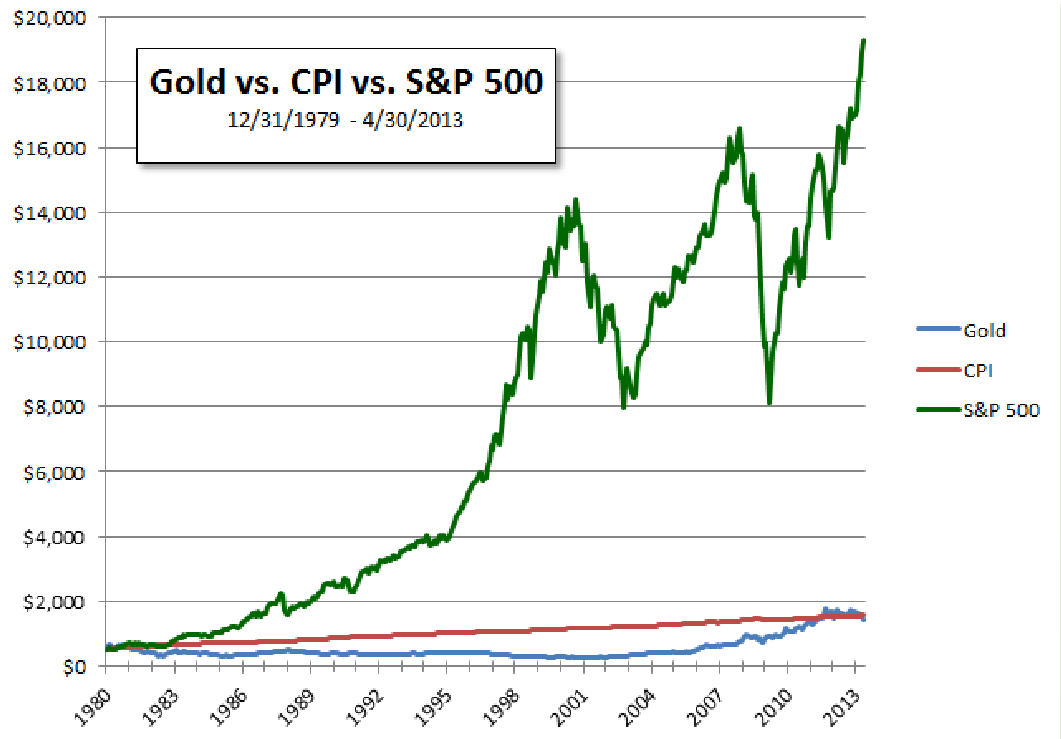
March 20, 2017

Agenda

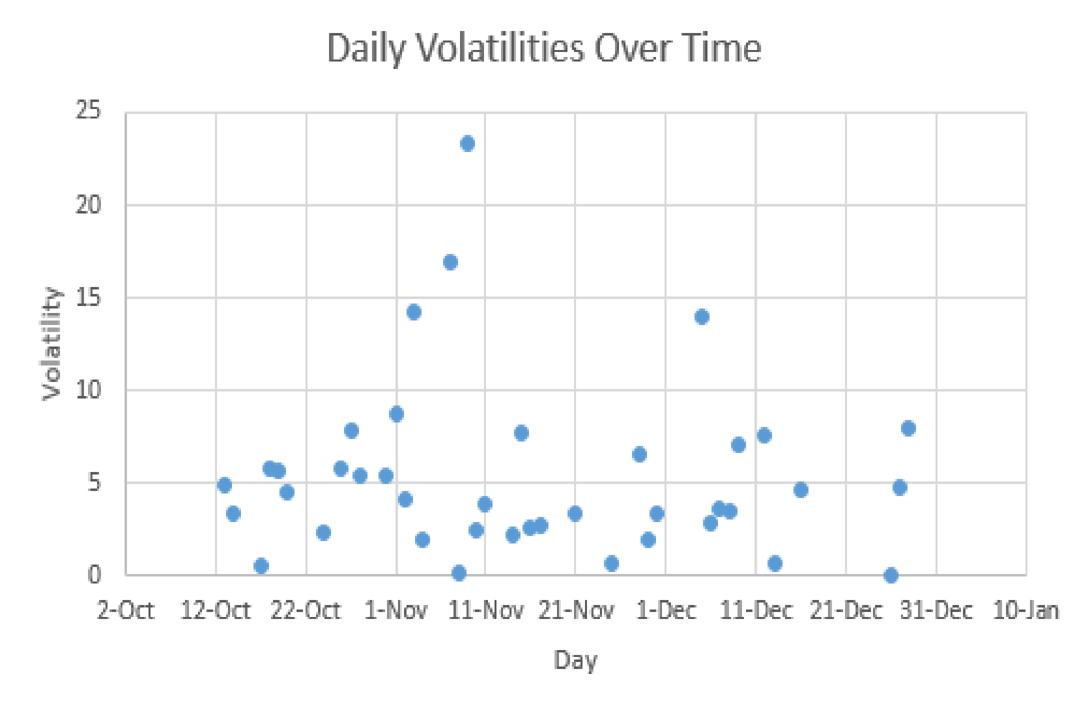
- Introduction
- Data
- Methods
- Results

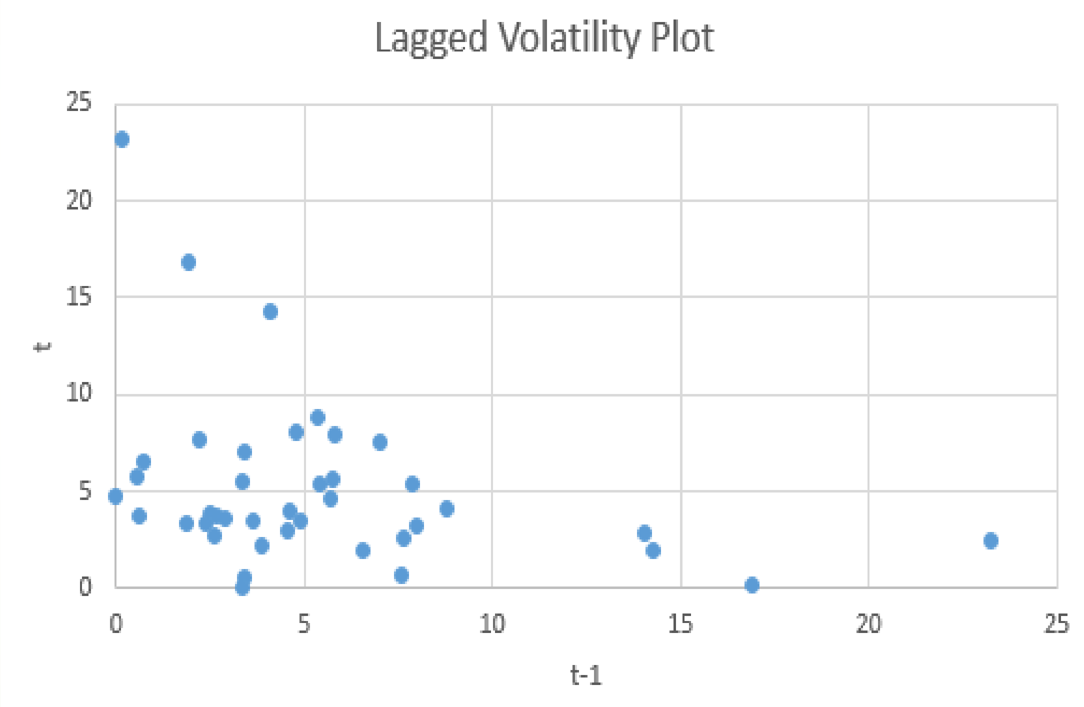
Introduction

- New York Stock Exchange (NYSE)
  - May 1792
- Standard & Poor's 500 (S&P 500)
  - March 1957
- Daily Volatility



Plots of Daily Volatilities



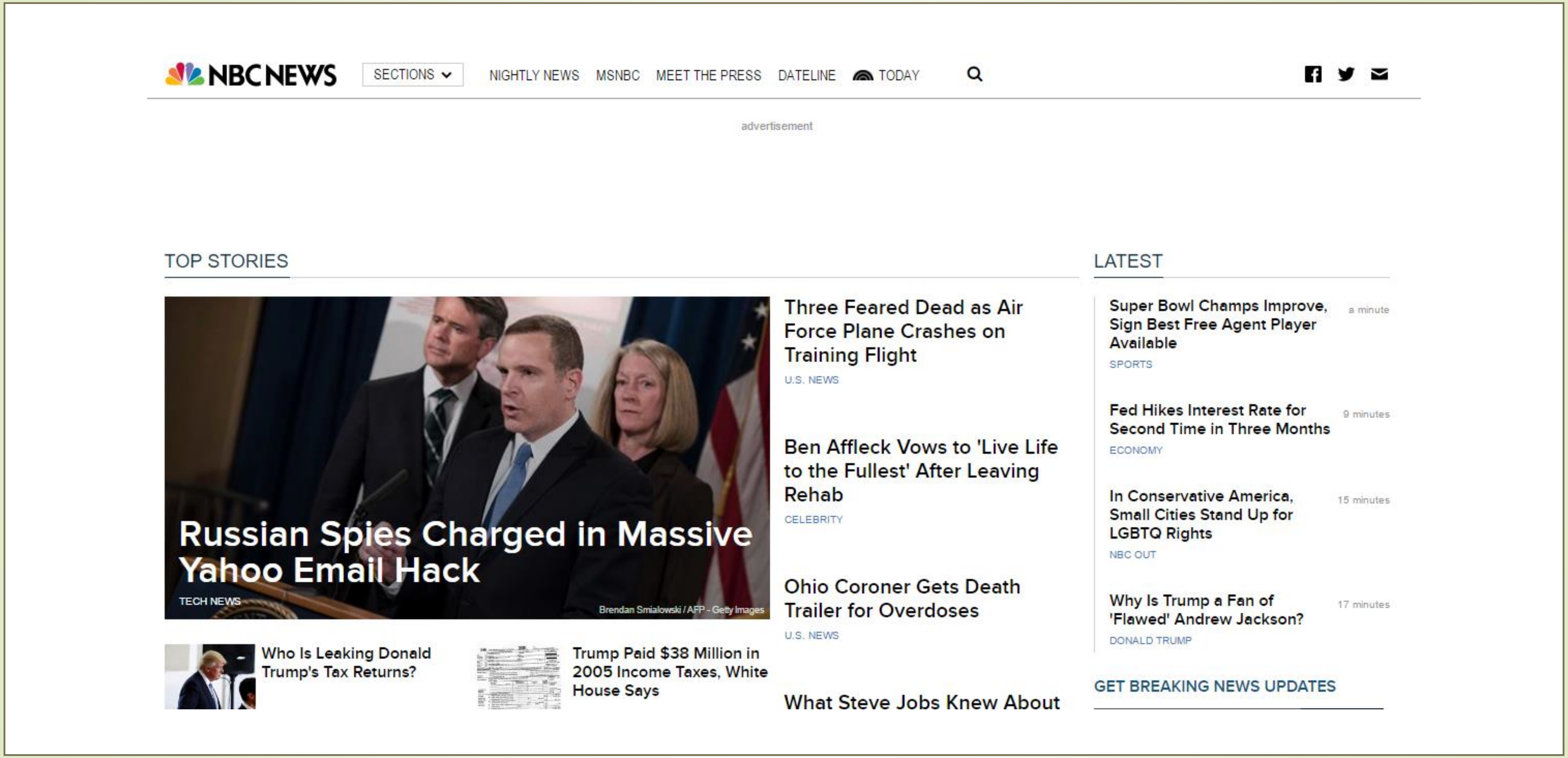


Natural Language Processing

- Method of manipulating text into a data set
  - Use text to make decisions
- Development since the 1930s
  - Dictionary Application: Birkbeck College, 1948
  - Code Breaking in WWII

Data

- October 13, 2016 – December 28, 2016
  - NBC News: Stock Market, Personal Finance, Economics, Politics
- Python Web Scraper
  - Beautiful Soup
  - URLLIB2
- NLP manipulation
  - NLTK



TOP STORIES

Russian Spies Charged in Massive Yahoo Email Hack

TECH NEWS

Brendan Smialowski / AFP - Getty Images

Three Feared Dead as Air Force Plane Crashes on Training Flight

U.S. NEWS

Ben Affleck Vows to 'Live Life to the Fullest' After Leaving Rehab

CELEBRITY

Ohio Coroner Gets Death Trailer for Overdoses

U.S. NEWS

What Steve Jobs Knew About

Who Is Leaking Donald Trump's Tax Returns?

Trump Paid $38 Million in 2005 Income Taxes, White House Says

LATEST

Super Bowl Champs Improve, Sign Best Free Agent Player Available — a minute

SPORTS

Fed Hikes Interest Rate for Second Time in Three Months — 9 minutes

ECONOMY

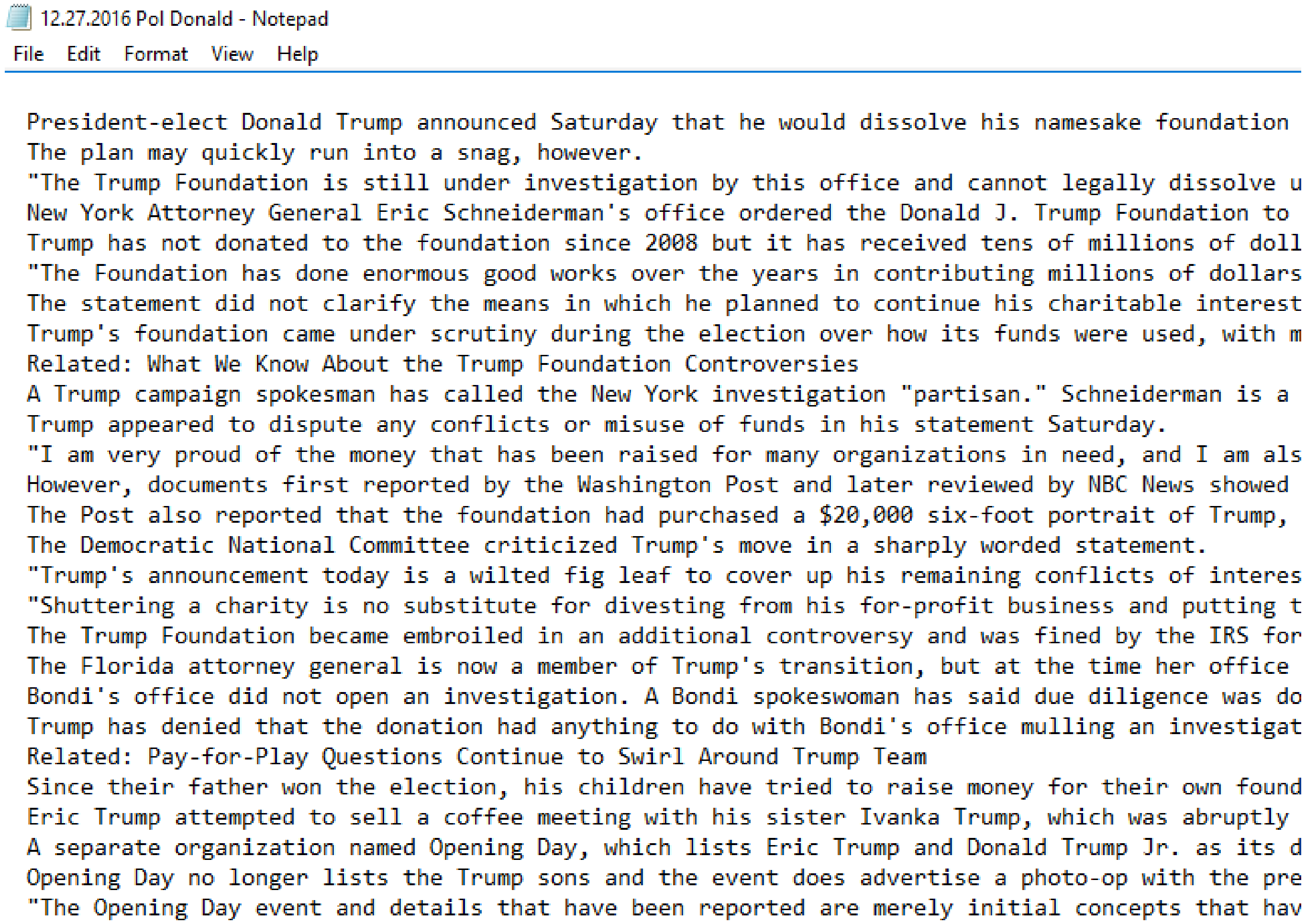
In Conservative America, Small Cities Stand Up for LGBTQ Rights — 15 minutes

NBC OUT

Why Is Trump a Fan of 'Flawed' Andrew Jackson? — 17 minutes

DONALD TRUMP

GET BREAKING NEWS UPDATES

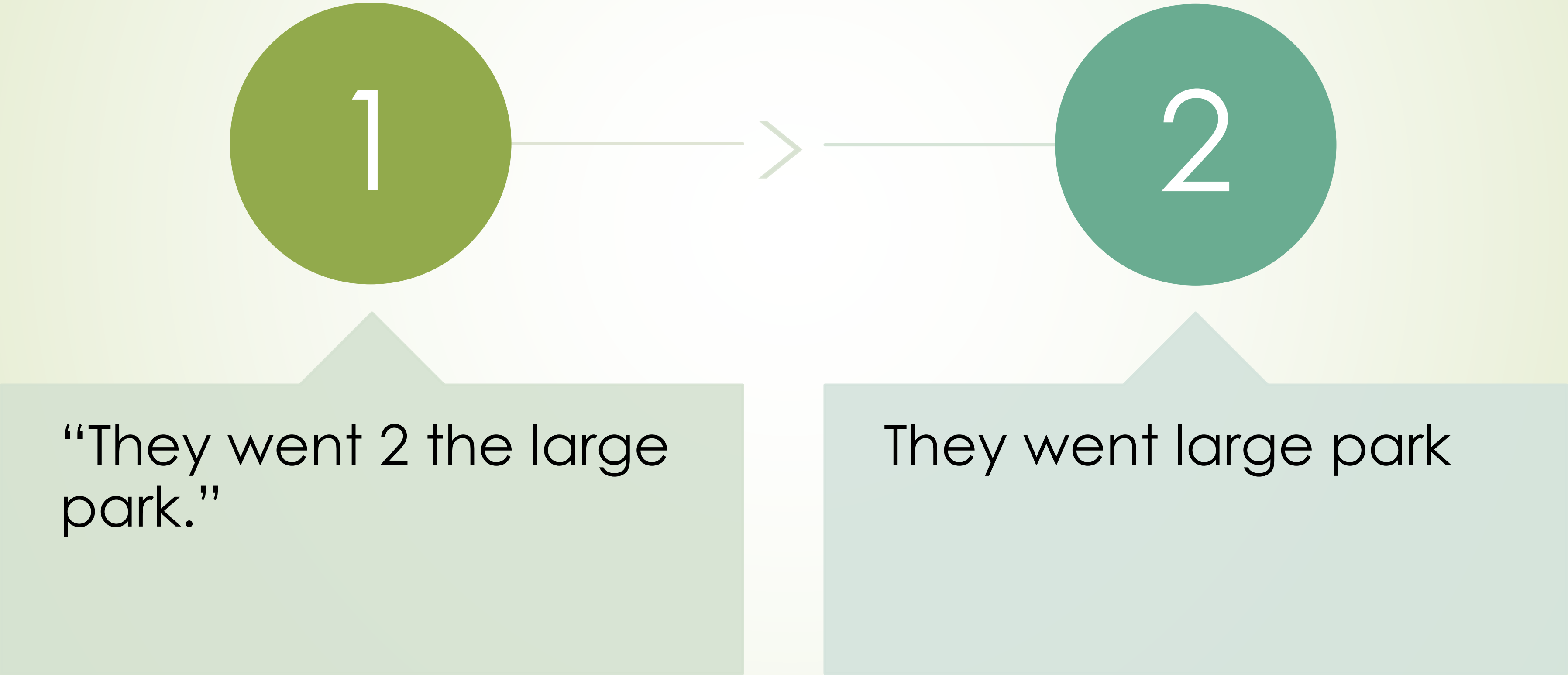


President-elect Donald Trump announced Saturday that he would dissolve his namesake foundation
The plan may quickly run into a snag, however.
"The Trump Foundation is still under investigation by this office and cannot legally dissolve u
New York Attorney General Eric Schneiderman's office ordered the Donald J. Trump Foundation to
Trump has not donated to the foundation since 2008 but it has received tens of millions of doll
"The Foundation has done enormous good works over the years in contributing millions of dollars
The statement did not clarify the means in which he planned to continue his charitable interest
Trump's foundation came under scrutiny during the election over how its funds were used, with m
Related: What We Know About the Trump Foundation Controversies
A Trump campaign spokesman has called the New York investigation "partisan." Schneiderman is a
Trump appeared to dispute any conflicts or misuse of funds in his statement Saturday.
"I am very proud of the money that has been raised for many organizations in need, and I am als
However, documents first reported by the Washington Post and later reviewed by NBC News showed
The Post also reported that the foundation had purchased a $20,000 six-foot portrait of Trump,
The Democratic National Committee criticized Trump's move in a sharply worded statement.
"Trump's announcement today is a wilted fig leaf to cover up his remaining conflicts of interes
"Shuttering a charity is no substitute for divesting from his for-profit business and putting t
The Trump Foundation became embroiled in an additional controversy and was fined by the IRS for
The Florida attorney general is now a member of Trump's transition, but at the time her office
Bondi's office did not open an investigation. A Bondi spokeswoman has said due diligence was do
Trump has denied that the donation had anything to do with Bondi's office mulling an investigat
Related: Pay-for-Play Questions Continue to Swirl Around Trump Team
Since their father won the election, his children have tried to raise money for their own found
Eric Trump attempted to sell a coffee meeting with his sister Ivanka Trump, which was abruptly
A separate organization named Opening Day, which lists Eric Trump and Donald Trump Jr. as its d
Opening Day no longer lists the Trump sons and the event does advertise a photo-op with the pre
"The Opening Day event and details that have been reported are merely initial concepts that hav

Natural Language Processing Techniques

- Number Removal
- Punctuation Removal
- Stop Word Removal
- Tokenizing
- Part of Speech (POS) tagging
- Stemming/Lemmatizing
- Document-Term Matrix Construction

NLP Techniques: Number, Punctuation, Stop Word Removal



1

“They went 2 the large park.”

2

They went large park

NLP Techniques: Tokenizing, Part of Speech (POS) Tagging

TOKENIZING

- They are going park
  - They
  - went
  - large
  - park

POS TAGGING

- They are going park
  - They (noun)
  - went (verb)
  - large (adjective)
  - park (noun)

NLP Techniques: Stemming vs. Lemmatizing Examples

- Based on POS tagging

- Multiply
  - Stemming: multip
  - Lemmatizing: multiple

- Features
  - Stemming: featur
  - Lemmatizing: feature

- Transparent
  - Stemming: transp
  - Lemmatizing: transparent

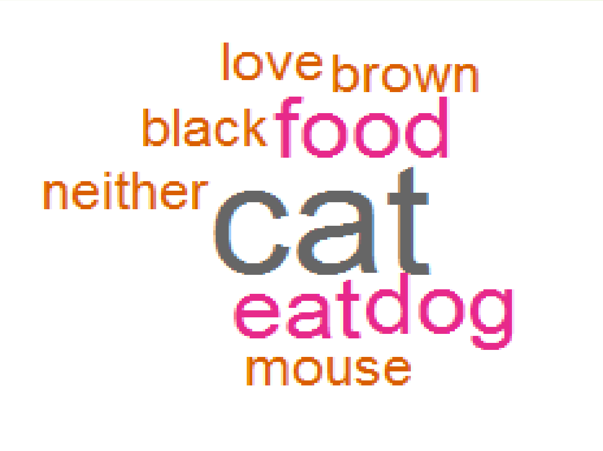
- Are
  - Stemming: ar
  - Lemmatizing: is

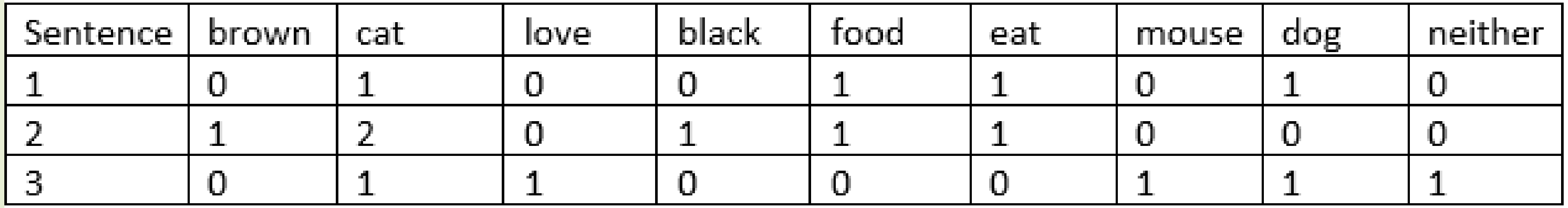
NLP Techniques: Document-Term Matrix (DTM)

- Frequency of each word in each article
- Word has to appear in at least 10% of articles
- 717 rows, 465 columns
- Low, Medium, High Volatilities

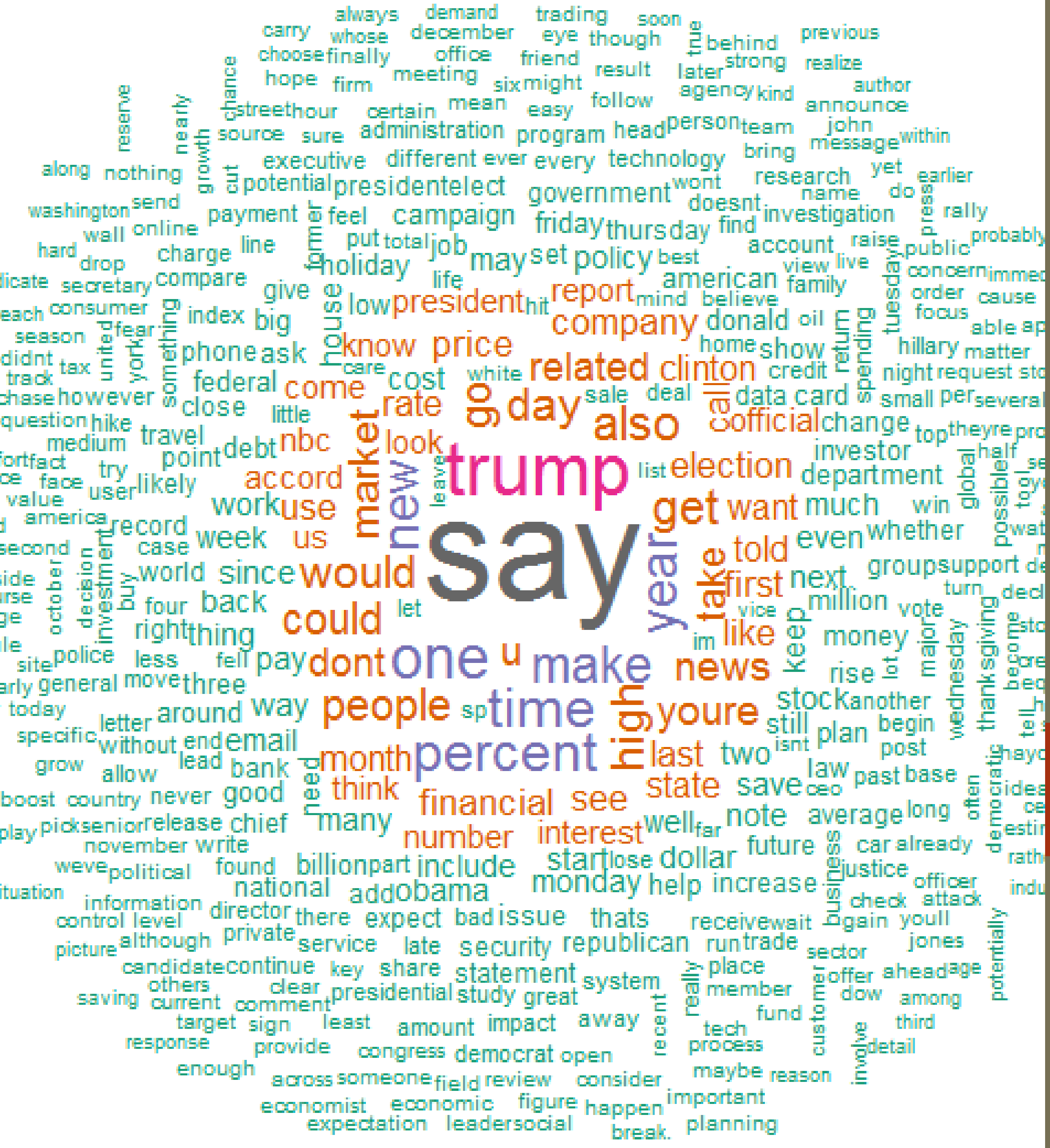
DTM Example

- Sentence 1: The cat and the dog eat their food.
- Sentence 2: Black cats and brown cats eat their food.
- Sentence 3: Neither cats nor dogs love mice.





Sentence	brown	cat	love	black	food	eat	mouse	dog	neither
1	0	1	0	0	1	1	0	1	0
2	1	2	0	1	1	1	0	0	0
3	0	1	1	0	0	0	1	1	1



Word Cloud for NBC News Articles

Methods

Boosting

xgboost

- Tree based
- Slow learning
- Weights on misclassified observations

Random Forests

ranger

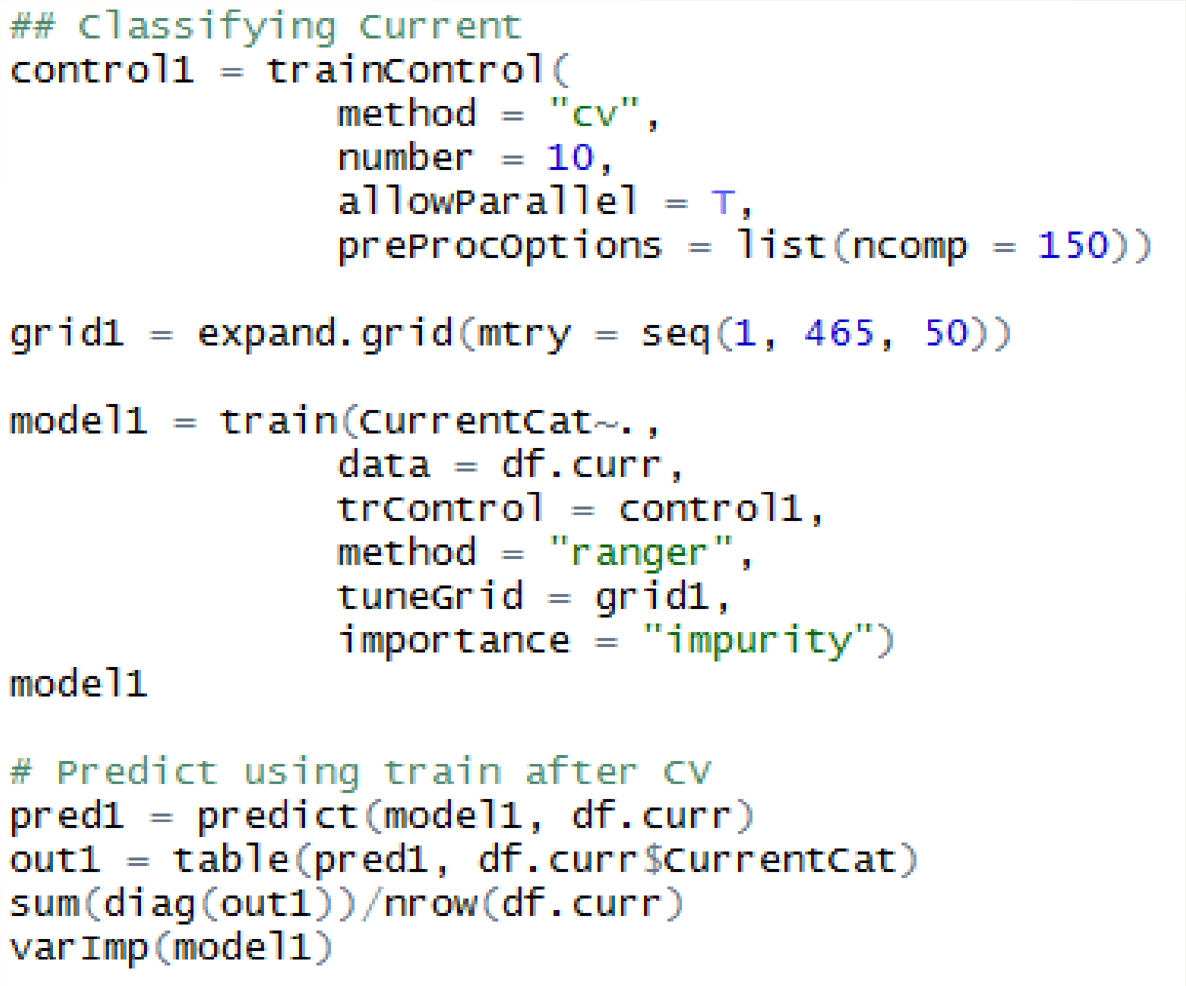
- Tree based
- Number of splits at each tree

Support Vector Machines

svmRadial

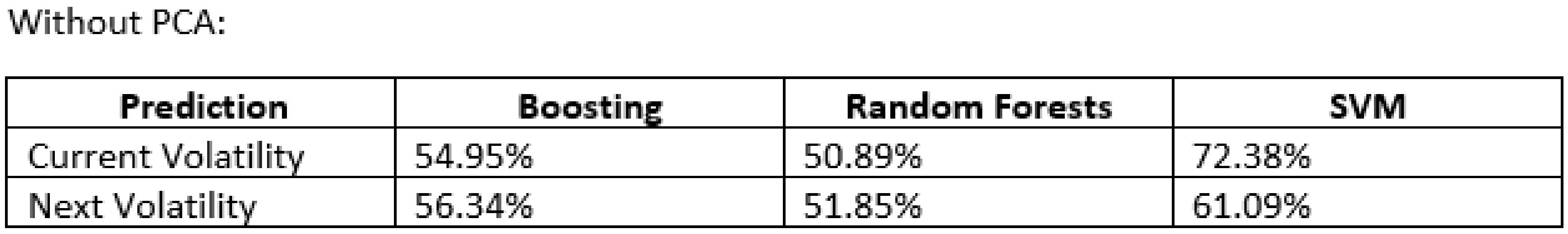
- State of the art classification method
- Nonlinear hyperplane to classify
- Specify width of error in hyperplane

Cross Validation



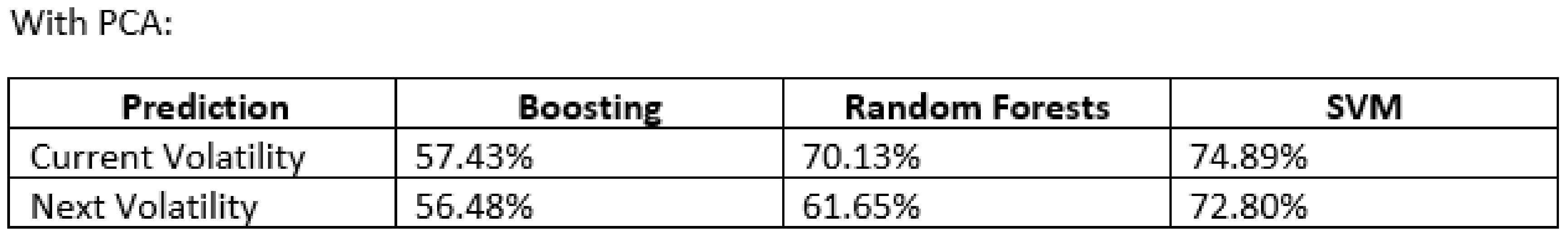
```
## Classifying Current
control1 = trainControl(
              method = "cv",
              number = 10,
              allowParallel = T,
              preProcOptions = list(ncomp = 150))

grid1 = expand.grid(mtry = seq(1, 465, 50))

model1 = train(CurrentCat~.,
              data = df.curr,
              trControl = control1,
              method = "ranger",
              tuneGrid = grid1,
              importance = "impurity")
model1

# Predict using train after CV
pred1 = predict(model1, df.curr)
out1 = table(pred1, df.curr$CurrentCat)
sum(diag(out1))/nrow(df.curr)
varImp(model1)
```

Results



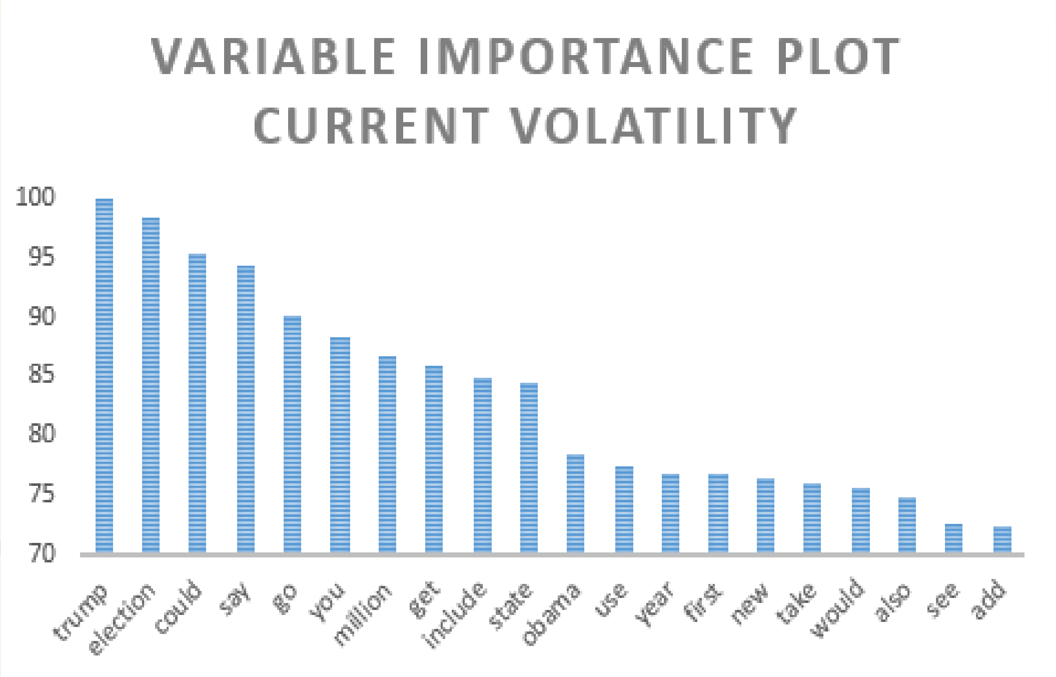
Without PCA:

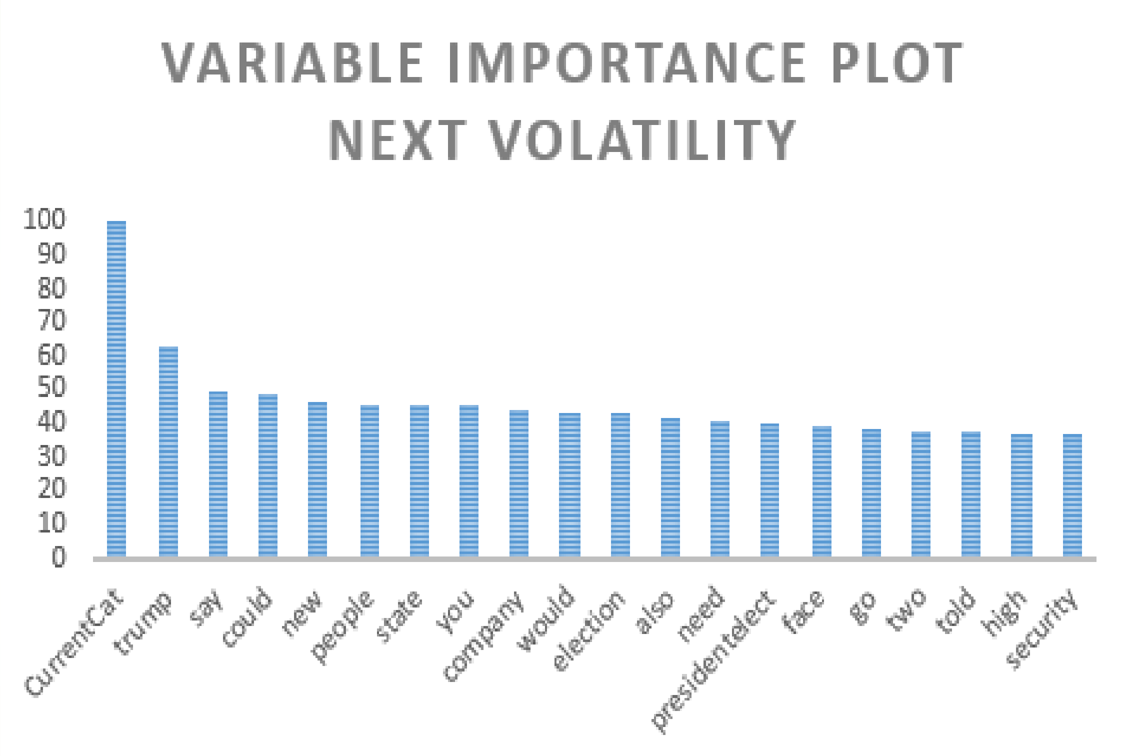
Prediction	Boosting	Random Forests	SVM
Current Volatility	54.95%	50.89%	72.38%
Next Volatility	56.34%	51.85%	61.09%



With PCA:

Prediction	Boosting	Random Forests	SVM
Current Volatility	57.43%	70.13%	74.89%
Next Volatility	56.48%	61.65%	72.80%


VARIABLE IMPORTANCE PLOT
CURRENT VOLATILITY
100
95
90
85
80
75
70
trump
election
could
say
go
you
million
get
include
state
obama
use
year
first
new
take
would
also
see
add


VARIABLE IMPORTANCE PLOT
NEXT VOLATILITY
100
90
80
70
60
50
40
30
20
10
0
CurrentCat
trump
say
could
new
people
state
you
company
would
election
also
need
presidentelect
face
go
two
told
high
security

Results

Conclusion

- Words showed predictive capability
- Wide variety of potential applications
- Learning new things and overcoming challenges

QUESTIONS