NLP – TWEETS FROM U.S. POLITICIANS

Suzanne McGann – Data Scientist

October 22, 2018

Agenda

- Capstone Topic Overview
- Why NLP?
- Data Wrangling
- Evaluate Models
- Next Steps
- Questions

Data

- FiveThirtyEight

- Tweets from…
  - *President Barack Obama*
  - *President Donald Trump*
  - *All Current U.S. Senators*

Questions

- Who published this tweet?
  - *President Obama or President Trump?*
  - *Republican or Democrat?*
- Which words were most predictive?

Why NLP?

- Written and verbal communication are incredibly important
  - *Failure to communicate often leads to delayed or poorly implemented projects*
  - *Communication break downs occur when we disregard others, and what they have to say*

Data Cleaning – Step 1

- Concatenate data frames
- Dropped '***bioguide_id***'
  - ***'user'*** field works as unique identifier
- Added '***party***' to presidential data

Data Cleaning - Step 2

- REGEX
  - *‰ÛÒ*
  - *&*
  - *ÏÓ•üè*

```
import re
```

REGULAR EXPRESSION

31 matches, 1126 steps (~4ms)

r" ([^0-9A-z\r:.'\/@\s]) " gm

TEST STR G

SWITCH TO UNIT TESTS

We%ÛªV protected more than 265 million acres...

Proud f these McKinley Tech students%ÛÓ>inspir...

On Int rnational Women,Ä∞√õ¬™s Day, @MichelleObama and I are inspi d by all of you who embrace your power to drive chang,Ä∞√õ_ https://t.co/Er9mIQlmgr

Chuck Berry rolled over everyone who came before him ,Äì and turned up everyone who came after. We'll miss you, Chuck. Be good.

Well said, Jimmy. That's exactly why we fought so hard for the ACA, and why we need to protect it for kids like Bil,Ä∞√õ_ https://t.co/jB3LXT940k

EXPLANATION

" ([^0-9A-z\r:.'\/@\s]) " gm

1st Capturing Group ([^0-9A-z\r:.'\/@\s])

Match a single character not present in the list below

[^0-9A-z\r:.'\/@\s]

0-9 a single character in the range between 0 (index 48) and 9 (index 57) (case sensitive)

A-z a single character in the range between A (index 65) and z (index 122) (case sensitive)

MATCH INFORMATION

Match 1

Full match 2-3 `%`

Group 1. 2-3 `%`

Match 2

Full match 3-4 `Û`

Group 1. 3-4 `Û`

Match 3

QUICK REFERENCE

Search reference

A single character of: a [abc]

https://regex101.com/

Data Cleaning – Step 3

```
def _removeNonAscii(s):
    my_str=''
    for i in s:
        if ord(i)<128:
            my_str=my_str+i
        else:
            my_str=my_str+'-'
    return my_str
```

```
for i in range(0,df_all.shape[0]):
    if len(re.findall(r'([^0-9A-z\r:.\'\/@\s])', df_all[['text']].iloc[i,0]) )>0:
        df_all.iloc[i,1] = _removeNonAscii(df_all[['text']].iloc[i,0])
```

Modeling -
Trump or Obama?

- Count Vectorizer + Multinomial Naïve Bayes
  - *Train Accuracy = .9728*
  - *Test Accuracy = .9651*

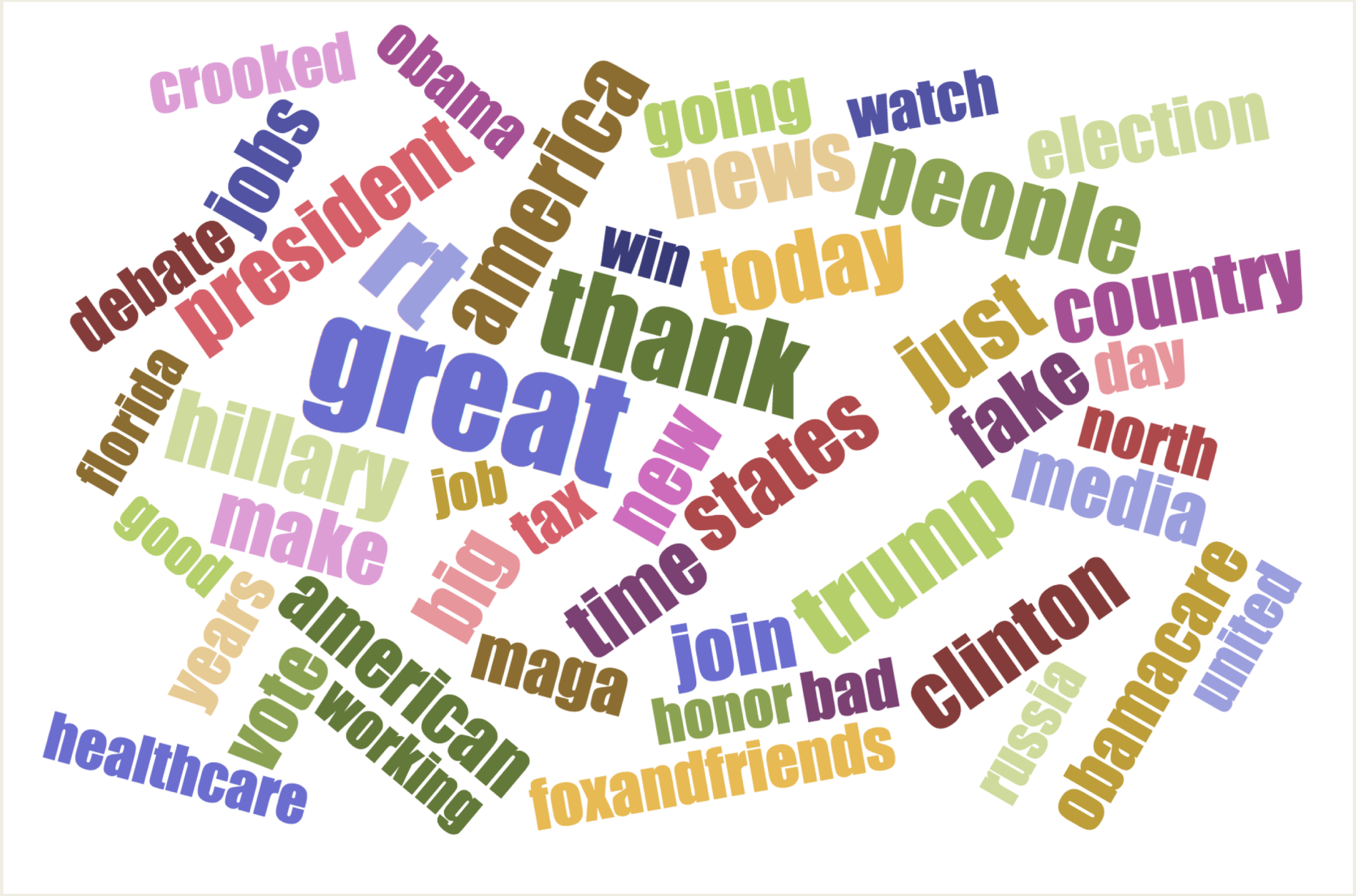
- Count Vectorizer + Random Forest
  - *Train Accuracy = .9588*
  - *Test Accuracy = .9589*

Modeling -
Trump or Obama?

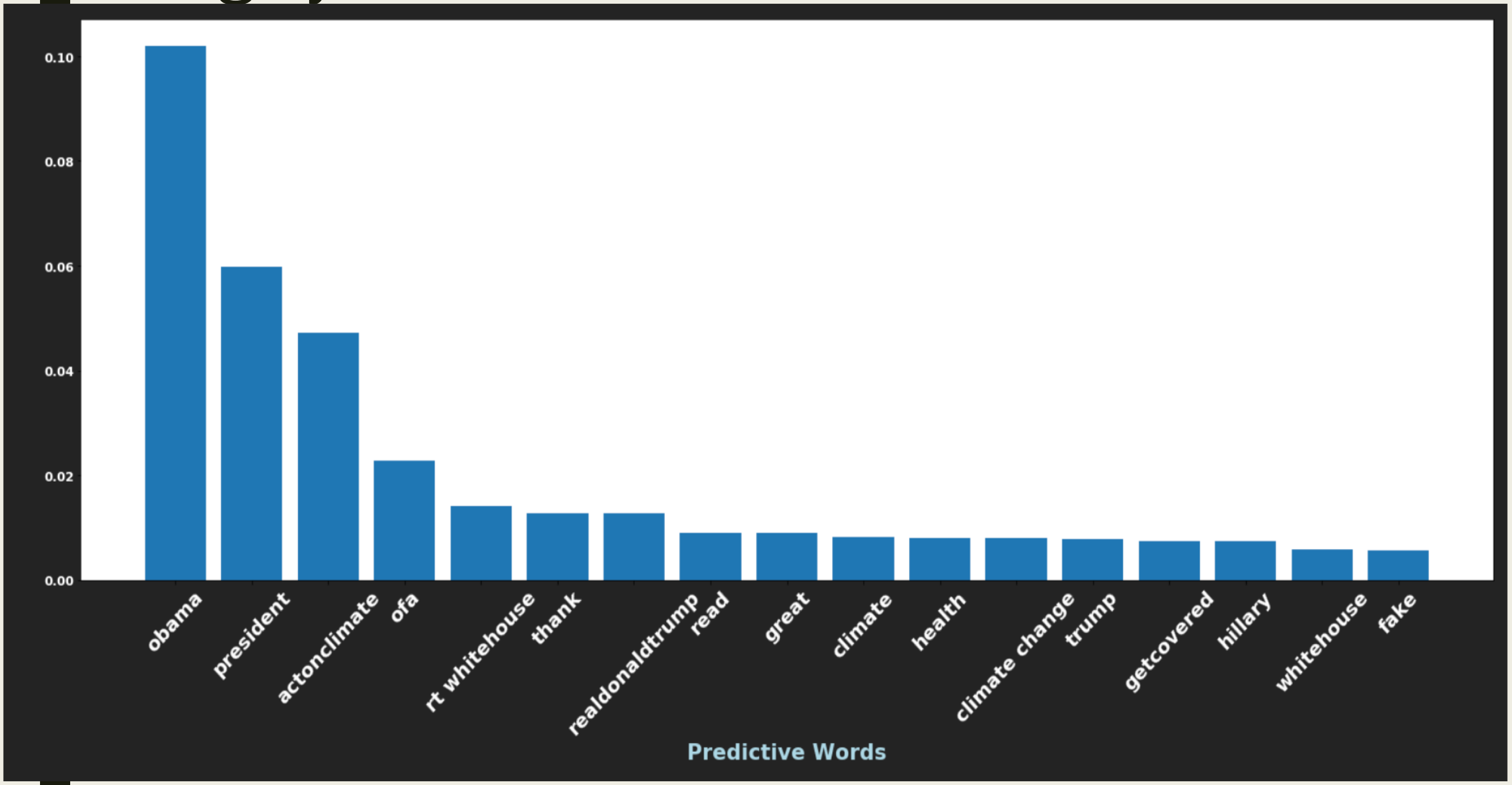
President	Precision	Recall	F1-Score	# of Tweets
Trump	0.96	0.97	0.97	647
Obama	0.97	0.96	0.96	641
Avg / Total	0.97	0.97	0.97	1288

Obama Word Cloud

Trump Word Cloud



Highly Predictive Words

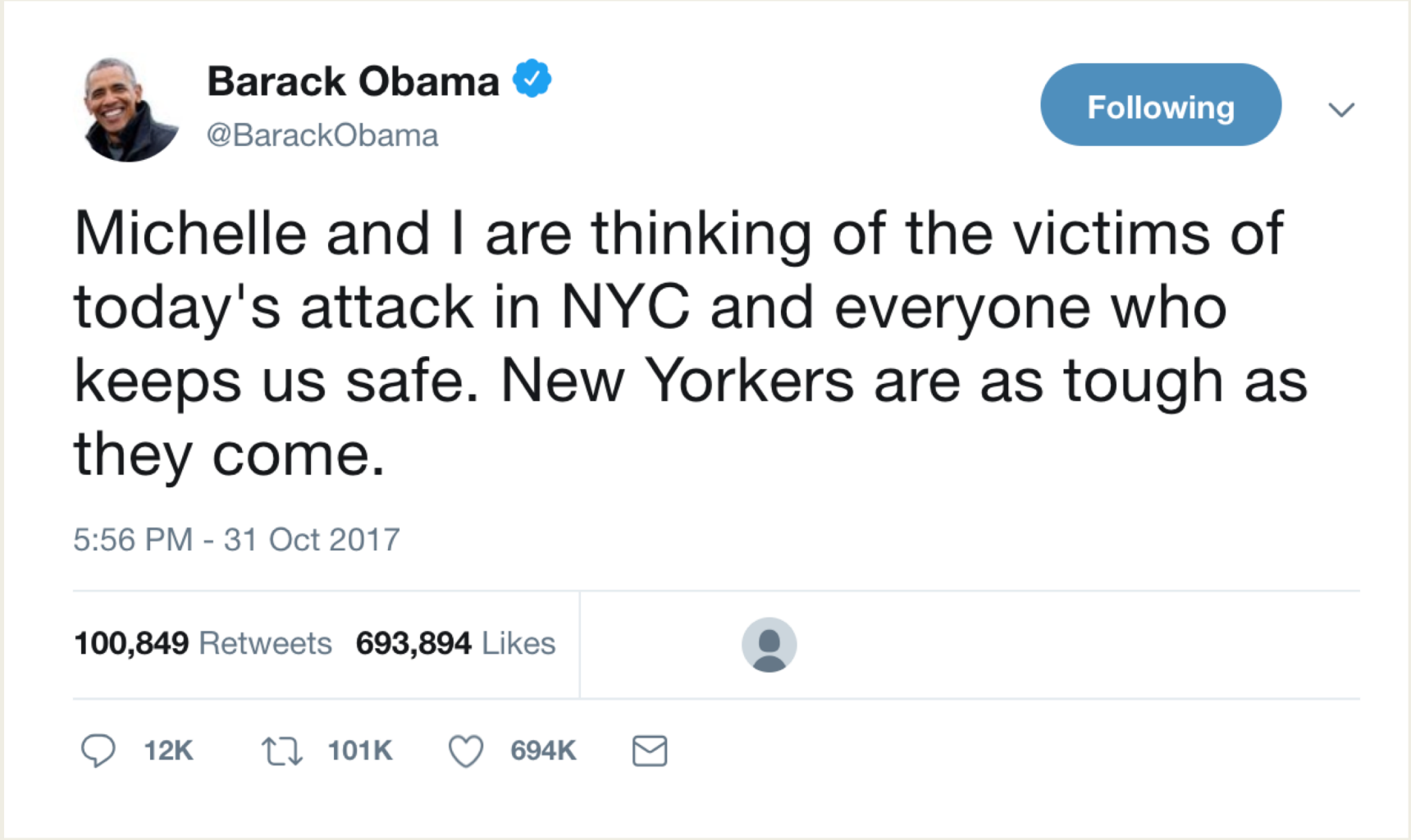


Misclassification Example 1

Misclassification Example 2

Misclassification Example 3

Misclassification Example 4



Modeling -
Republican or Democrat?

- Count Vectorizer + Multinomial Naïve Bayes
  - *Train Accuracy = .8404*
  - *Test Accuracy = .8479*
- Count Vectorizer + Random Forest
  - *Train Accuracy = .7841*
  - *Test Accuracy = .7937*
- Count Vectorizer + Logistic Regression
  - *Train Accuracy = .8612*
  - *Test Accuracy = .8733*

Modeling -
Republican or Democrat?

Party	Precision	Recall	F1-Score	# of Tweets
Republican	0.87	0.89	0.88	29,902
Democrat	0.88	0.85	0.87	27,858
Avg / Total	0.87	0.87	0.87	57,760

Next Steps

- Senate Level Analysis
  - *Top words by state, by party*
  - *Top words by region, by party*

- Sentiment Analysis

NLP – TWEETS FROM U.S. POLITICIANS

Suzanne McGann – Data Scientist

October 22, 2018