

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

- %oÜÖ
- &
- ö•üè

```
import re
```

REGULAR EXPRESSION

31 matches, 1126 steps (~4ms)

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

" gm

TEST STRINGS

SWITCH TO UNIT TESTS ▾

We're protected more than 265 million acres...

Proud of these McKinley Tech students who inspire...

On International Women's Day, @MichelleObama and I are inspired by all of you who embrace your power to drive change.

<https://t.co/Er9mIQlmgm>

Chuck Berry rolled over everyone who came before him 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 Bill Clinton.

<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

w

[^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

Table

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'([^\u00a1-\u00ff]+)', 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
<i>Trump</i>	0.96	0.97	0.97	647
<i>Obama</i>	0.97	0.96	0.96	641
<i>Avg / Total</i>	0.97	0.97	0.97	1288

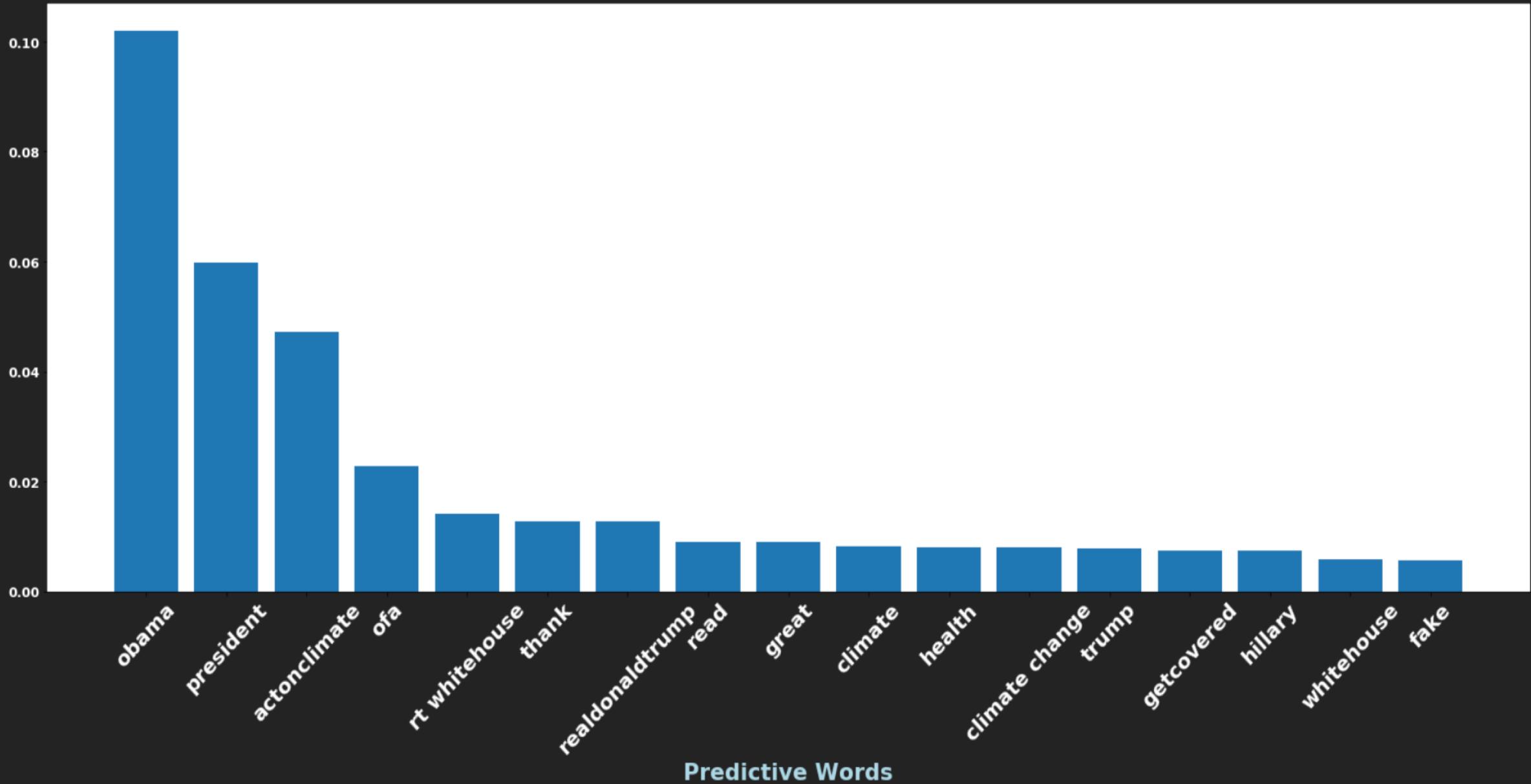
Obama Word Cloud



Trump Word Cloud



Highly Predictive Words



Misclassification Example 1

 **Barack Obama** 
@BarackObama

 **Following** 

I'm grateful to [@SenJohnMcCain](#) for his lifetime of service to our country. Congratulations, John, on receiving this year's Liberty Medal.

3:59 PM - 16 Oct 2017

89,705 Retweets 665,376 Likes



 18K  90K  665K 

Misclassification Example 2

 **Donald J. Trump** 
@realDonaldTrump

The world is noticing, thanks!

Iana del fenty @glamourizes
Replying to @realDonaldTrump

Trump is a winner! No matter what the haters say, he's actually making this country great again! Thank you Mr. President for actually caring

3:44 AM - 20 Sep 2017

11,441 Retweets 55,521 Likes



 8.8K  11K  56K 

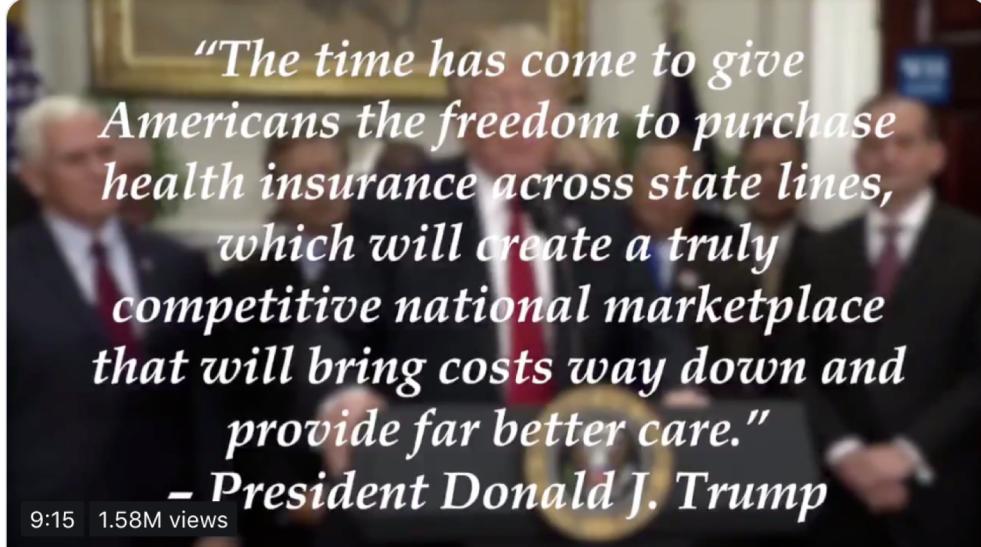
Misclassification Example 3

 Donald J. Trump 
@realDonaldTrump

Follow ▾

The time has come to take action to
IMPROVE access, INCREASE choices, and
LOWER COSTS for HEALTHCARE!

→ 45.wh.gov/Sp9y4H



*"The time has come to give
Americans the freedom to purchase
health insurance across state lines,
which will create a truly
competitive national marketplace
that will bring costs way down and
provide far better care."*

- President Donald J. Trump

9:15 | 1.58M views

PRESIDENT TRUMP SIGNS HEALTHCARE EXECUTIVE ORDER

The time has come to give Americans the freedom to purchase health ins across state lines - creating a truly competitive national marketplace that will bring costs way down & provide far better care.

Misclassification Example 4

Barack Obama

@BarackObama

Following

Michelle and I are thinking of the victims of today's attack in NYC and everyone who keeps us safe. New Yorkers are as tough as they come.

5:56 PM - 31 Oct 2017

100,849 Retweets 693,894 Likes

12K 101K 694K

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
<i>Republican</i>	0.87	0.89	0.88	29,902
<i>Democrat</i>	0.88	0.85	0.87	27,858
<i>Avg / Total</i>	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