

Tracking Conspiracies on Reddit during the 2016 US Election

Jared Delora-Ellefson
Data Scientist



AGENDA

Problem Statement 01

02 Collecting/Processing the Data

The spaCy model 03

04 Analyzing the
Model Results

Future Steps 05

Problem Statement



Identify conspiracy trends in the political discussions occurring on Reddit between Jan 2016 and Dec 2016. This period covers the year 2016 of the US Presidential Election of Donald Trump.

The fundamental questions to be investigated:

- Did the conspiracies that plagued the 2016 US Presidential election begin with chatter on social media?
- How much of the spread of these conspiracies was due to political figures spreading misinformation?

Project Milestones

Pull Data:

- 2016 Reddit Data Pulled ✓
- Created small subsets of each month for model training (AWS set up and ready) ✓
- Twitter data for handles ✓

Analyze Reddit
Comments using
NLP Model on AWS ✓

Step 1

Step 2

Step 3

Step 4

Use NLP (spaCy) to Build
Entity Recognition Model ✓

Synthesize Results
In Process

Project Flow

It's been shown that Russia pushed the following conspiracies:

- Seth Rich
- Pizzagate + Podesta
- BlueLivesMatter
- HerEmails

Subreddits:

r/the_donald
r/Republican
r/Conservative
r/TedCruz
r/MarcoRubio

Twitter handles to investigate:

@realdonaldtrump, @realrogerstone1,
@donaldtrumpjr, @BreitbartNews,
@infowarsmedia, @zerohedge

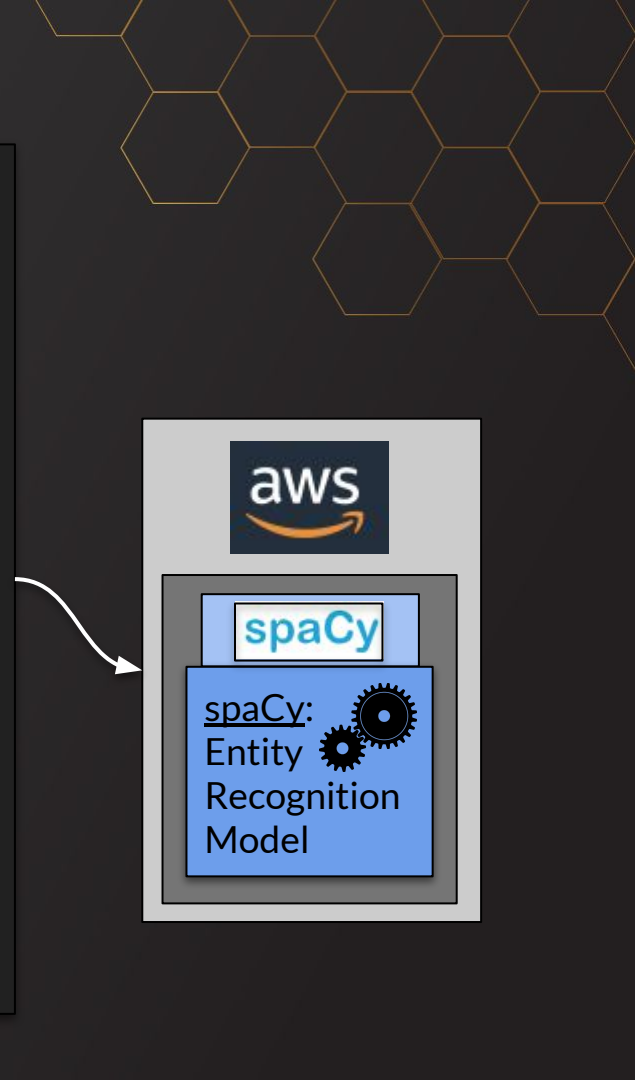
References:

Online Political Discourse in the Trump Era
RISHAB NITHYANAND, et al.

<https://arxiv.org/pdf/1711.05303.pdf>

Fivethirtyeight.com

<https://fivethirtyeight.com/features/dissecting-trumps-most-rabid-online-following/>



Data Processing



r/the_donald
Between 1-5 million
comments per month

Analyze all 2020 r/the_donald Reddit
Comments using NLP Model on AWS

> 20 Million
Comments

Filter on the Comments and
collect a Training set of
10,000 comments. Add 30%
comments that are not
conspiracy related.



Train spaCy
model with
Training set



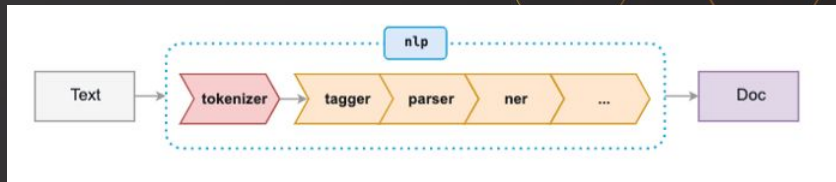
Process all 2016
comments using AWS

The comments for the year we're
filtered on a list of conspiracy terms

The spaCy model

The spaCy model used the following technologies:

- sense2vec
- tok2vec
- Prodigy for Entity Recognition Training



spaCy Model Metrics:

- Precision: ~91
- Recall: ~93
- F-Score: ~92

This model performs well. It has low bias and low variance.

Example Text Analysis:

```
Loading Model: donald_model...
The text we will be processing...
```

```
-----
George Soros went to the Seth Rich supermarket where he saw Black Lives Matter
protesting the killing of Seth Rich over Hillary's emails and Clinton's emails.
George smiled knowing that blm, Pizzagate, Blue Lives Matter are just Globalist plots to cover
up Clinton's emails and the human trafficking and pedo pizzagate activities taking place in Benghazi.
-----
```

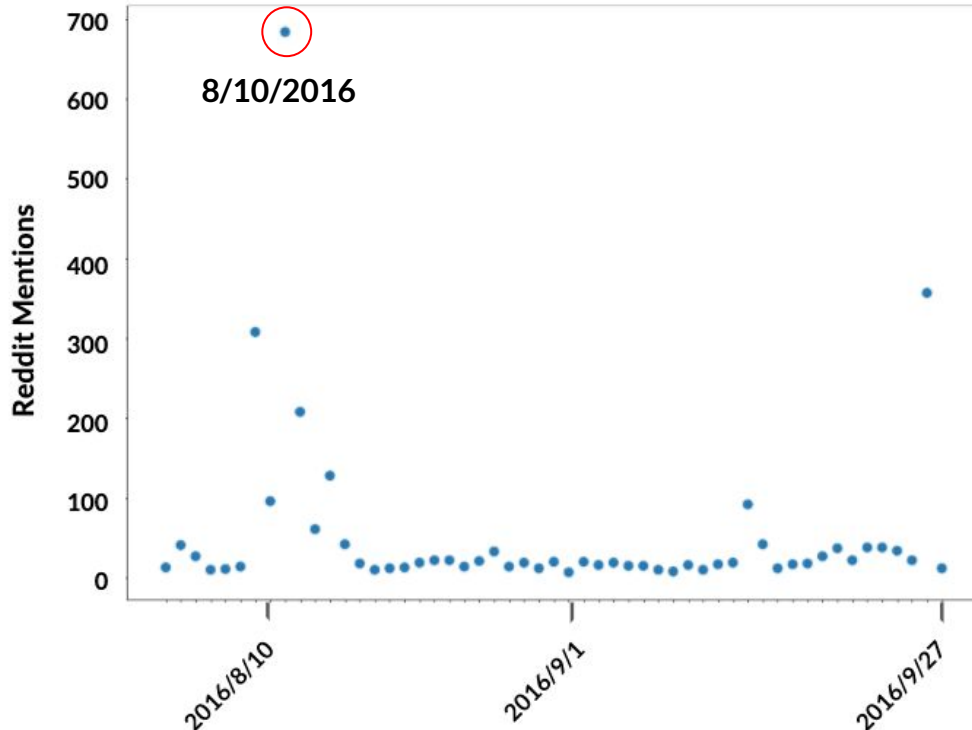
```
Entities detected:
```

```
{'Globalist', 'Blue Lives Matter', 'human trafficking', 'Pizzagate', 'Seth Rich', 'pedo', 'blm', 'George Soros', 'Benghazi', 'Black Lives Matter', 'pizzagate'}
```


Analyzing Model Results - Seth Rich Conspiracy


Mentions of Seth Rich in comments on r/the_donald

Seth Rich was Murdered on July 10, 2016



Next Steps



1. Complete Seth Rich Reddit Analysis.
 2. Complete Reddit Analysis for the other identified conspiracies.
 3. Collect and analyze the remaining subreddits and twitter accounts.
 4. Use the NER model to analyze the Reddit/Twitter accounts listed above.
 5. Exploratory Analysis and Correlation Findings
- 

THANKS



Do you have any questions?
JaredDelora@gmail.com

CREDITS: This presentation template was created
by **Slidesgo**, including icons by **Flaticon**, and
infographics & images by **Freepik**
Please keep this slide for attribution.