# Fake News:

## From Definition to Identification

Baurjan Safi

General Assembly, Data Science Immersive-5

September 2017

# Plan of Presentation

- “Fake News” Definition
- Data and EDA
- Machine Learning Results
- Conclusions and Recommendations
- Discussion

# Definition of Fake News

**Fake news**

A type of journalism or propaganda that consists of deliberate misinformation or hoaxes spread via traditional print and broadcast news media or online social media.

Fake news is written and published with the intent to mislead in order to gain financially or politically, often with sensationalist, exaggerated, or patently false headlines that grab attention

Wikipedia

# Data and EDA

- Initial dataset of fake news found on Kaggle.com
  - 12,999 observation over the past year.
  - Over 200 news resources from a wide specter of political affiliations.
- DataCamp.com article by Katharine Jarmul (Berlin)
  - 6,000 observations (3k – Real, 3K – Fake)
  - Unspecified sources
  - Mostly political articles
- "Real" news scraped with help of "newspaper" Python package
  - Over 6,000 observations
  - From 15 news sites, mostly with high rank of credibility

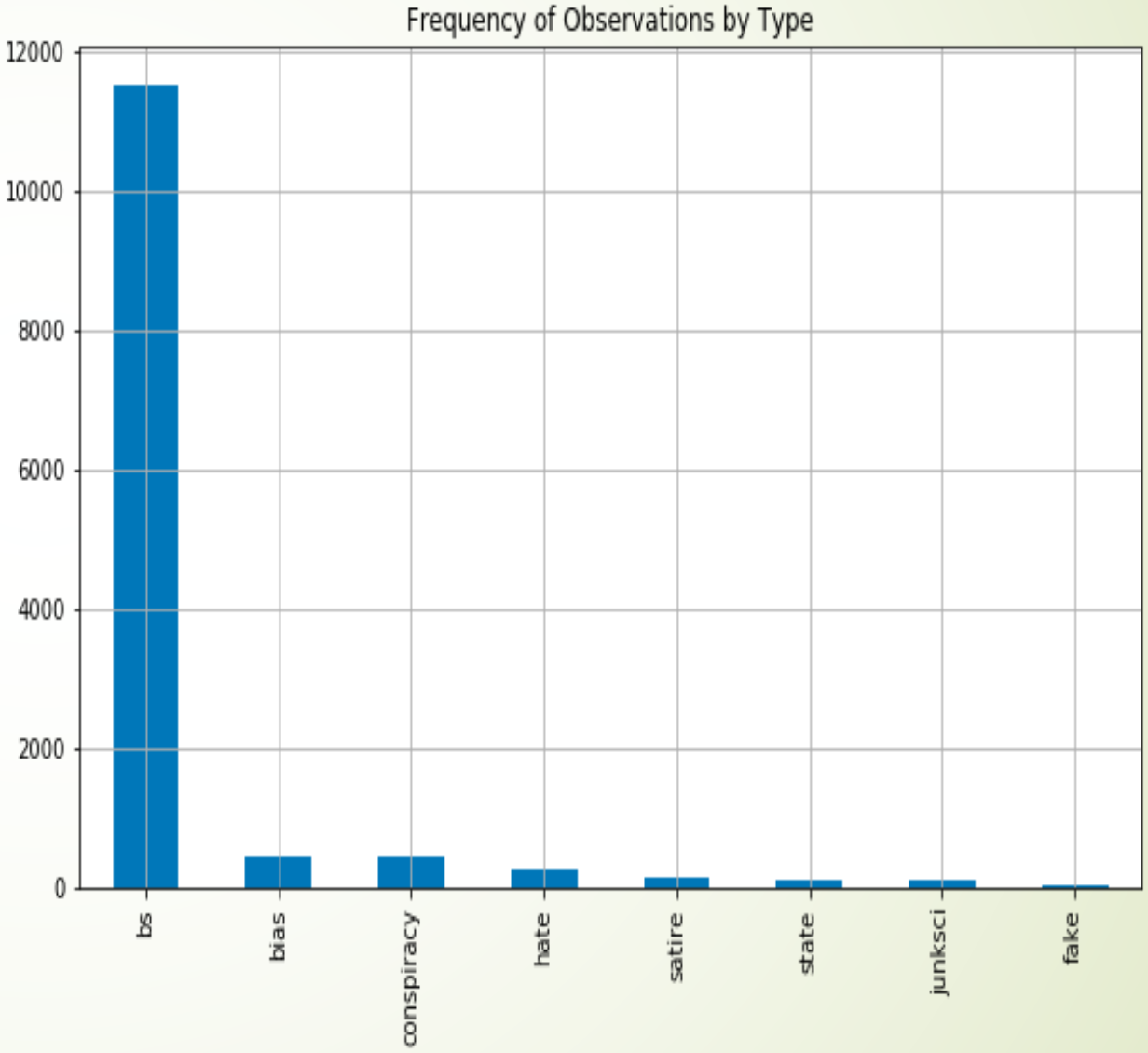
# EDA

Media Bias and Fact Check:

- Political affiliation:
  - Extreme Left
  - Left
  - Center Left
  - Center
  - Center Right
  - Right
  - Extreme Right
- Credibility of the news source:
  - High
  - Mixed
  - Low
- Conspiracy / Pseudoscience
- Satire

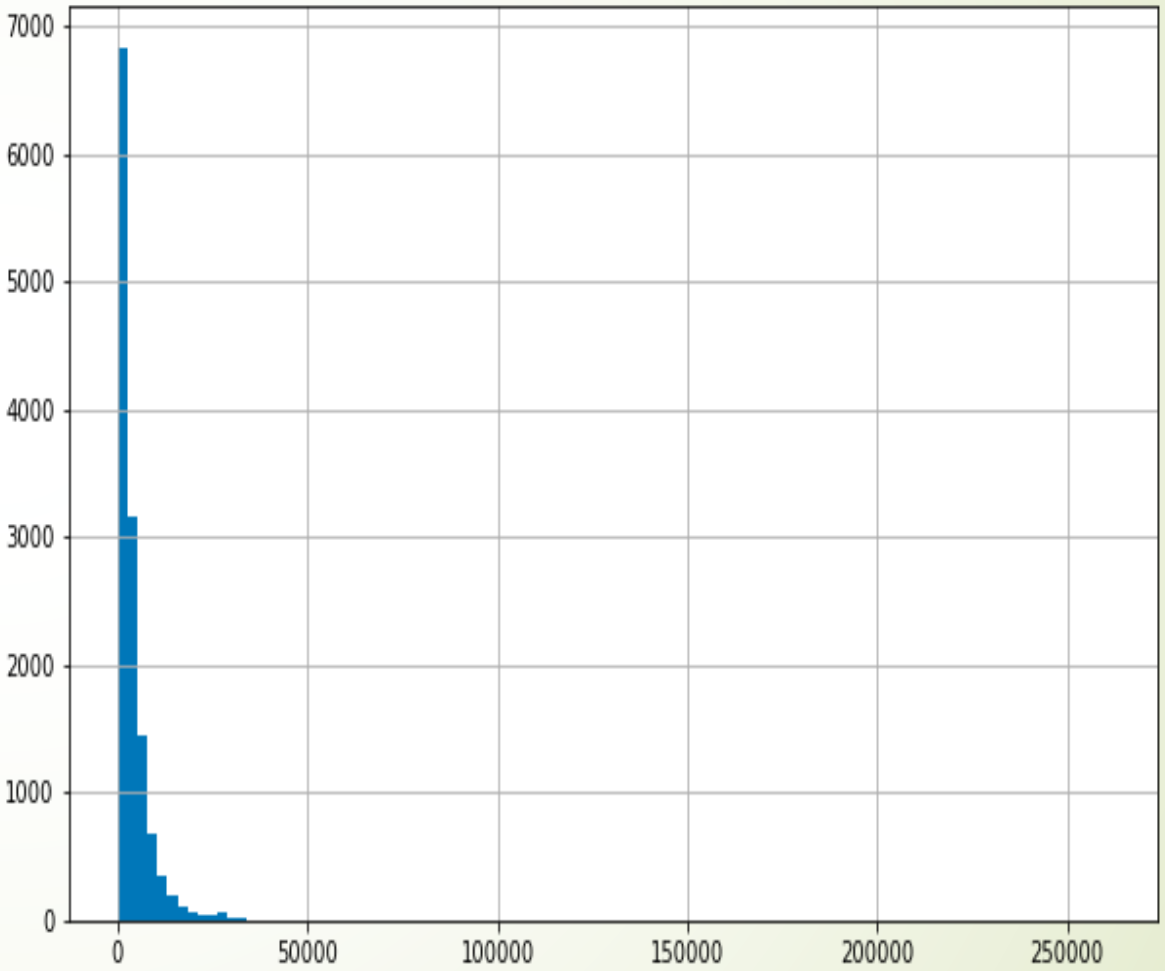
# Kaggle “Fake News” Dataset

| Classification | Number |
|---|---|
| bs | 11492 |
| bias | 443 |
| conspiracy | 430 |
| hate | 246 |
| satire | 146 |
| state | 121 |
| junksci | 102 |
| fake | 19 |

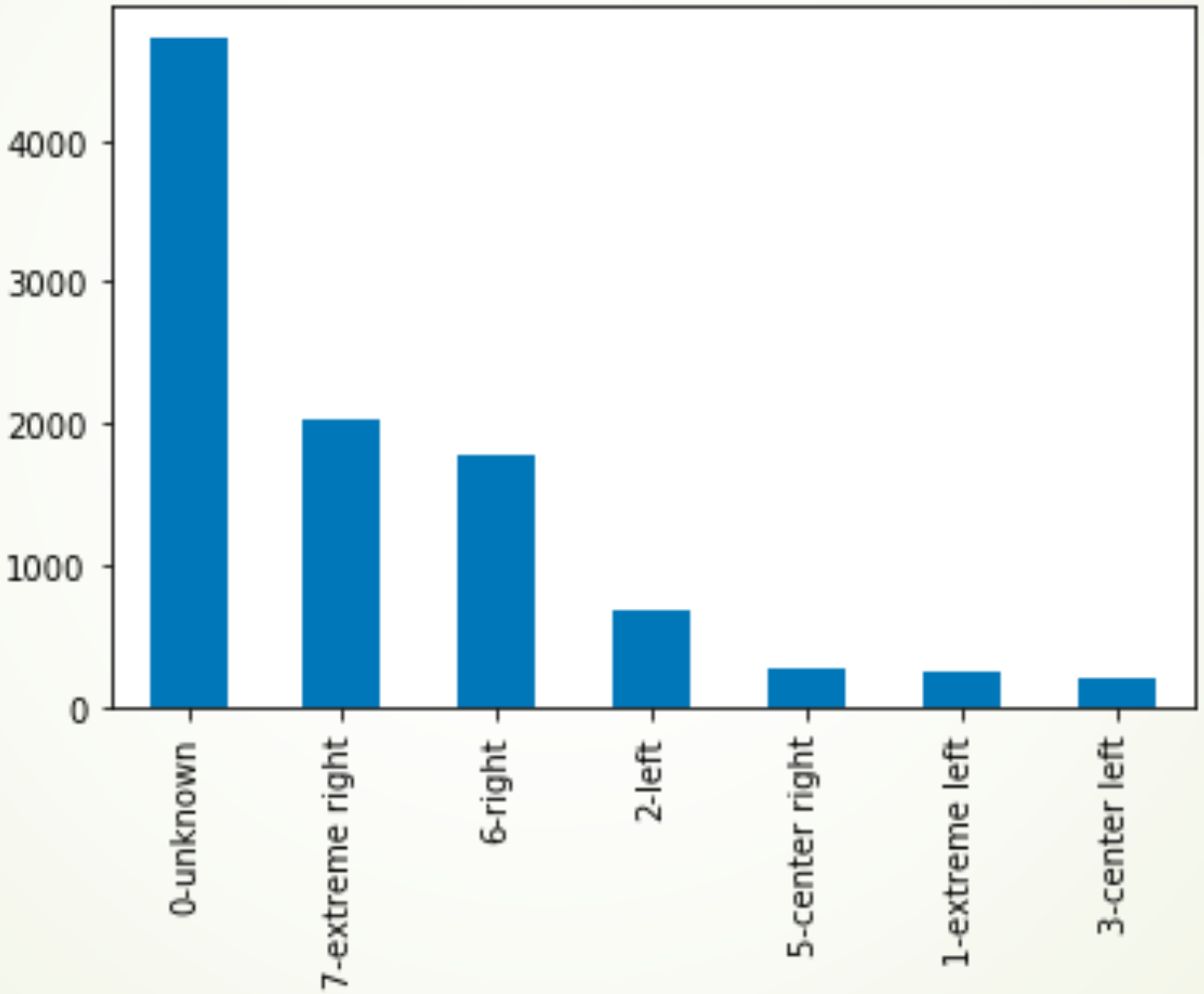


# Kaggle “Fake News” Dataset

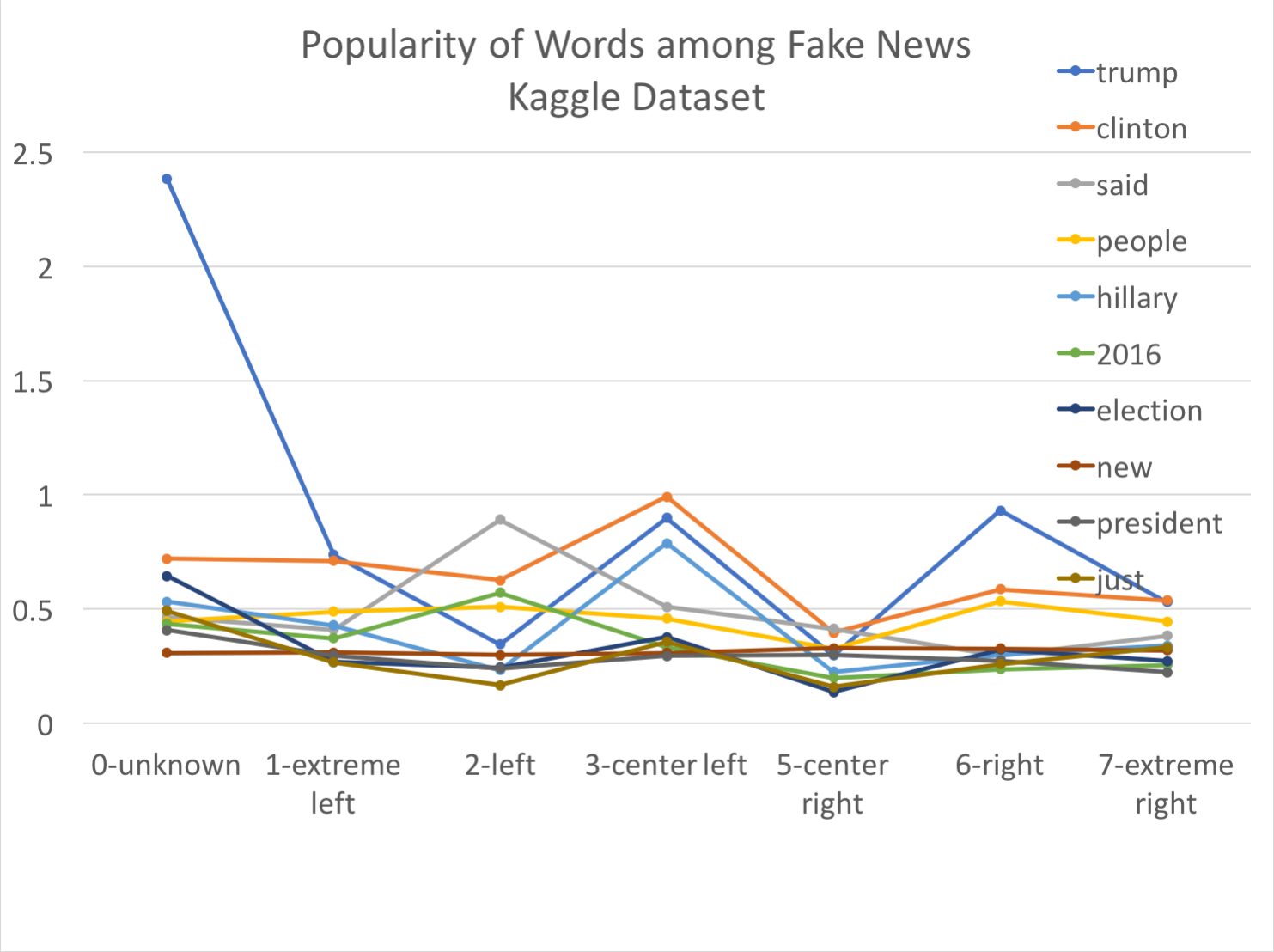
- Dropped:
  - Other languages other than English
  - Other classes than 'bs'
  - All texts with length of less than 500 and longer than 12,000 signs
- Ran it through all 215 news sources and determined about 140 on them on the political affiliation specter.



# News Spread on a Political Specter



# The list of the popular words across media

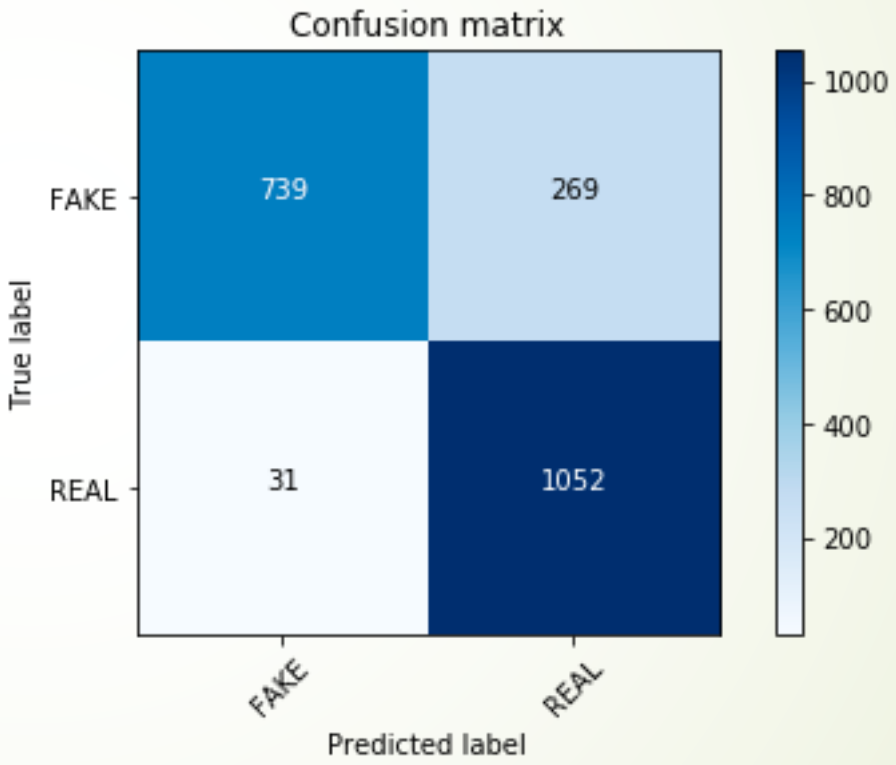


# Methodology

- Build Model with Multinomial Naïve Bayes on DataCamp's dataset
- Transform the Kaggle's fake news dataset
- Transform the scraped supposedly real news dataset
- Merge the Kaggle and scraped datasets and refit the model on Multinomial NB
- Compare other classification methods and compare the results

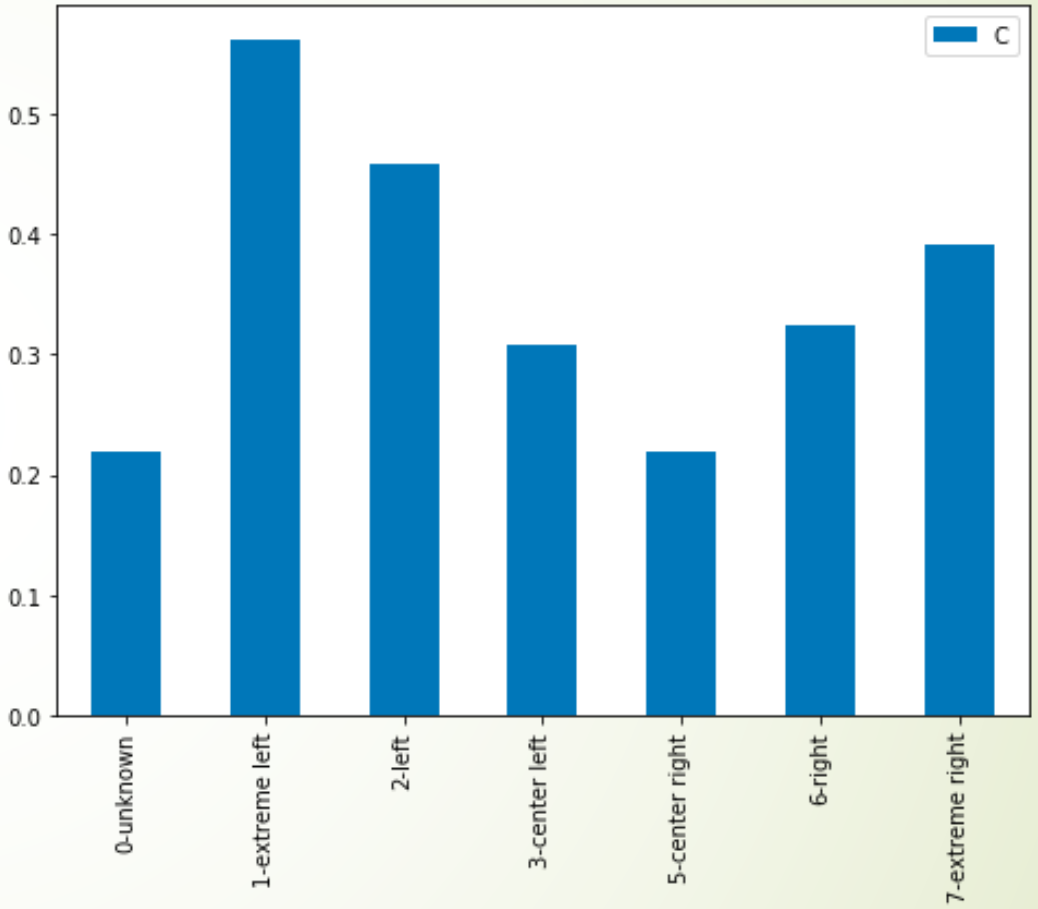
# Results of Datacamp Dataset on Multinomial Naïve Bayes

Accuracy - 0.857



# “Credibility” of the “Fake News”

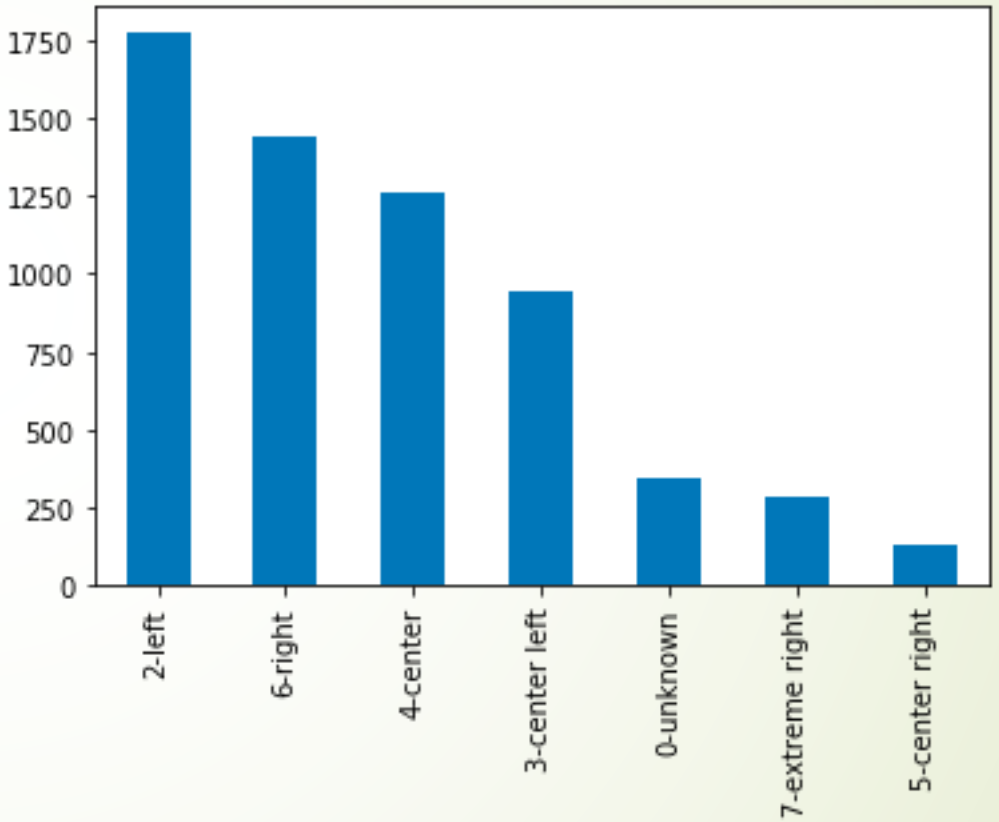
| | FAKE | REAL | TOTAL | CREDIBILITY |
|---|---|---|---|---|
| **0-unknown** | 3681 | 1037 | 4718 | 0.219797 |
| **1-extreme left** | 109 | 140 | 249 | 0.562249 |
| **2-left** | 367 | 310 | 677 | 0.457903 |
| **3-center left** | 141 | 63 | 204 | 0.308824 |
| **5-center right** | 213 | 60 | 273 | 0.219780 |
| **6-right** | 1197 | 573 | 1770 | 0.323729 |
| **7-extreme right** | 1239 | 795 | 2034 | 0.390855 |



# Web Scraped “Real” News Cross-cut

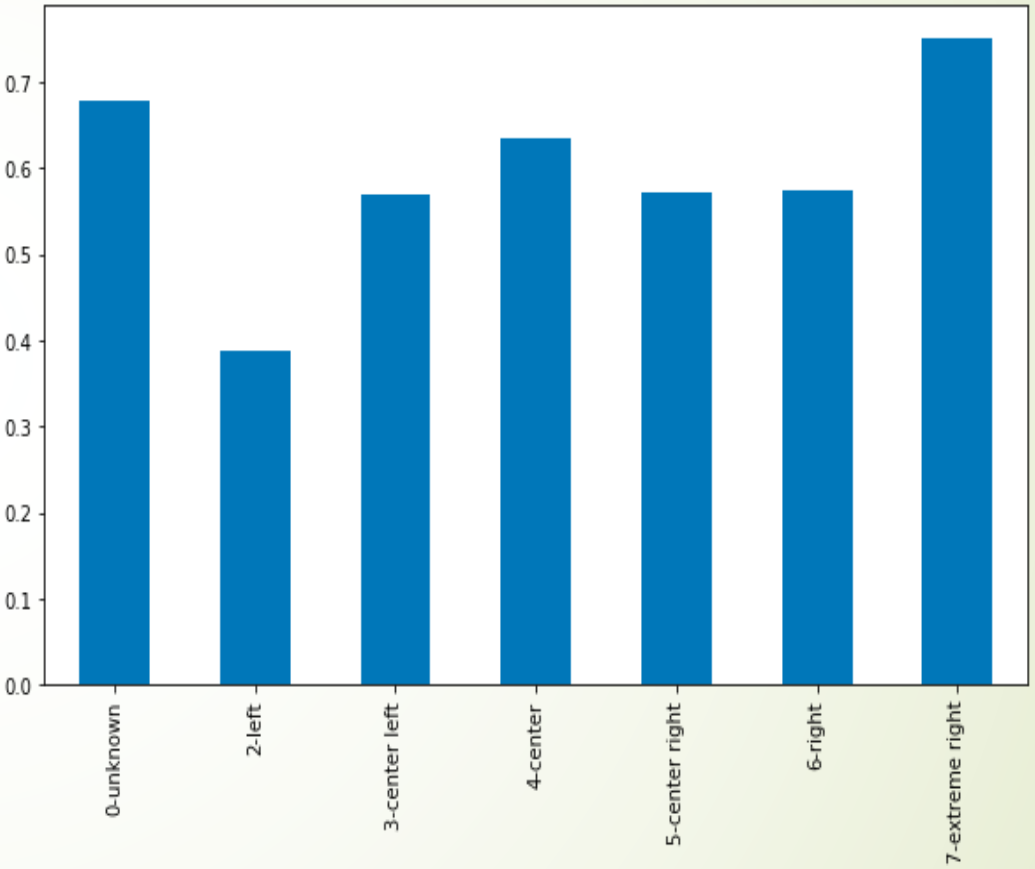
| Party Affiliation | # of News |
|---|---|
| 0-unknown | 341 |
| 2-left | 1774 |
| 3-center left | 944 |
| 4-center | 1260 |
| 5-center right | 126 |
| 6-right | 1444 |
| 7-extreme right | 287 |

Number of ”Real” News / Affiliation



# ”Real” News on Multinomial Naïve Bayes

| | FAKE | REAL | TOTAL | CREDIBILITY |
|---|---|---|---|---|
| **0-unknown** | 105 | 220 | 325 | 0.676923 |
| **2-left** | 752 | 478 | 1230 | 0.388618 |
| **3-center left** | 403 | 534 | 937 | 0.569904 |
| **4-center** | 454 | 787 | 1241 | 0.634166 |
| **5-center right** | 54 | 72 | 126 | 0.571429 |
| **6-right** | 580 | 780 | 1360 | 0.573529 |
| **7-extreme right** | 71 | 215 | 286 | 0.751748 |

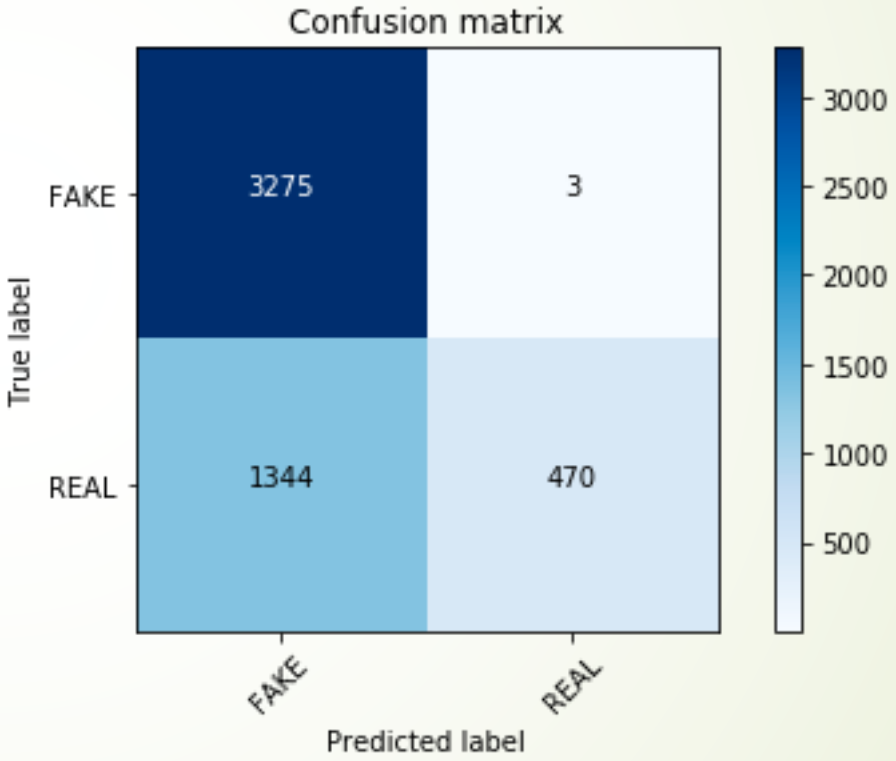


# Merged Dataset on Multinomial Naïve Bayes

Accuracy - 0.735

Fake news correctly predicted ~ 100%

Real news correctly predicted - 0.28%



# Other Algorithms

| Algorithm | Accuracy Score |
|---|---|
| LinearSVC() | 0.907 |
| RandomForestClassifier() | 0.909 |
| GaussianNB() | 0.828 |
| GradientBoostingClassifier() | 0.904 |

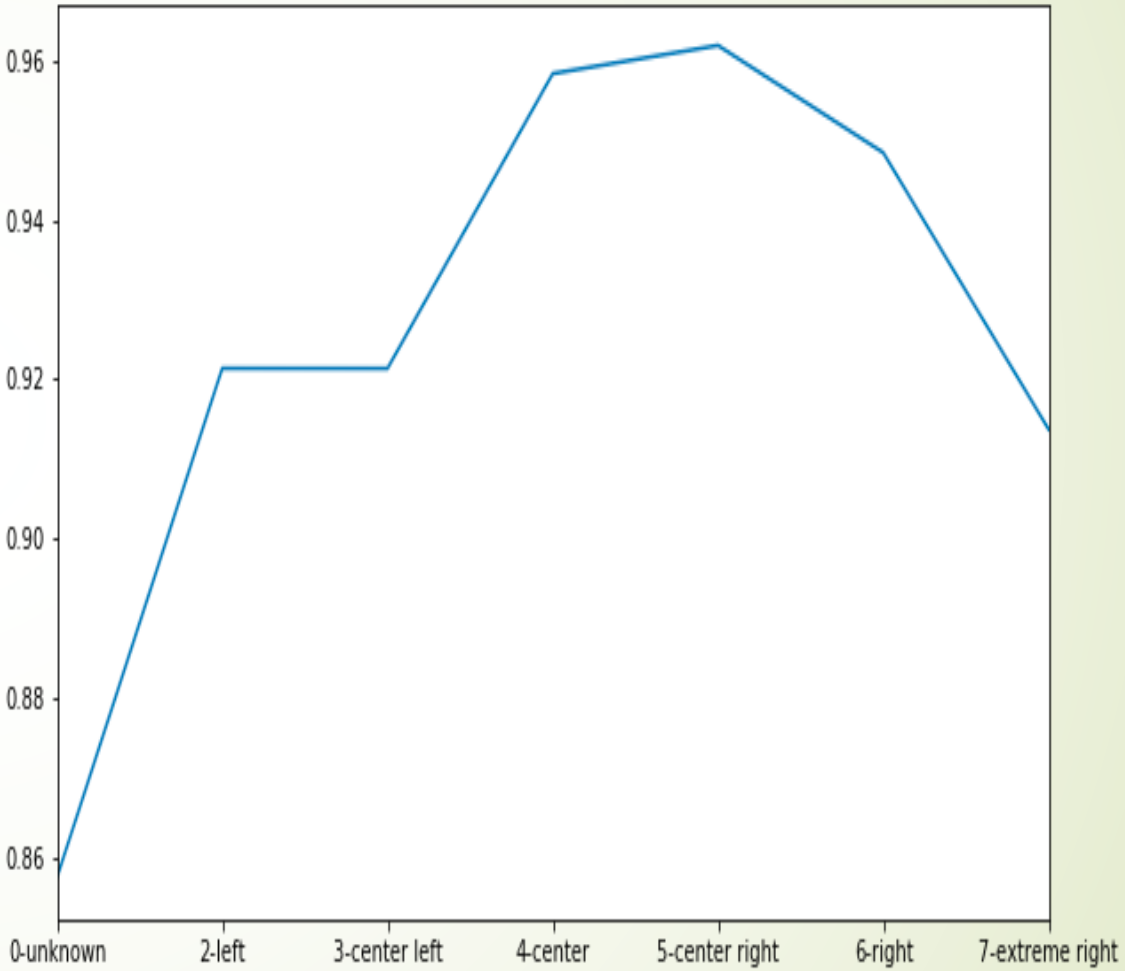
# Random Forest

| Correctly predicted | |
|---|---|
| Fake news | 99% |
| Real news | 93% |

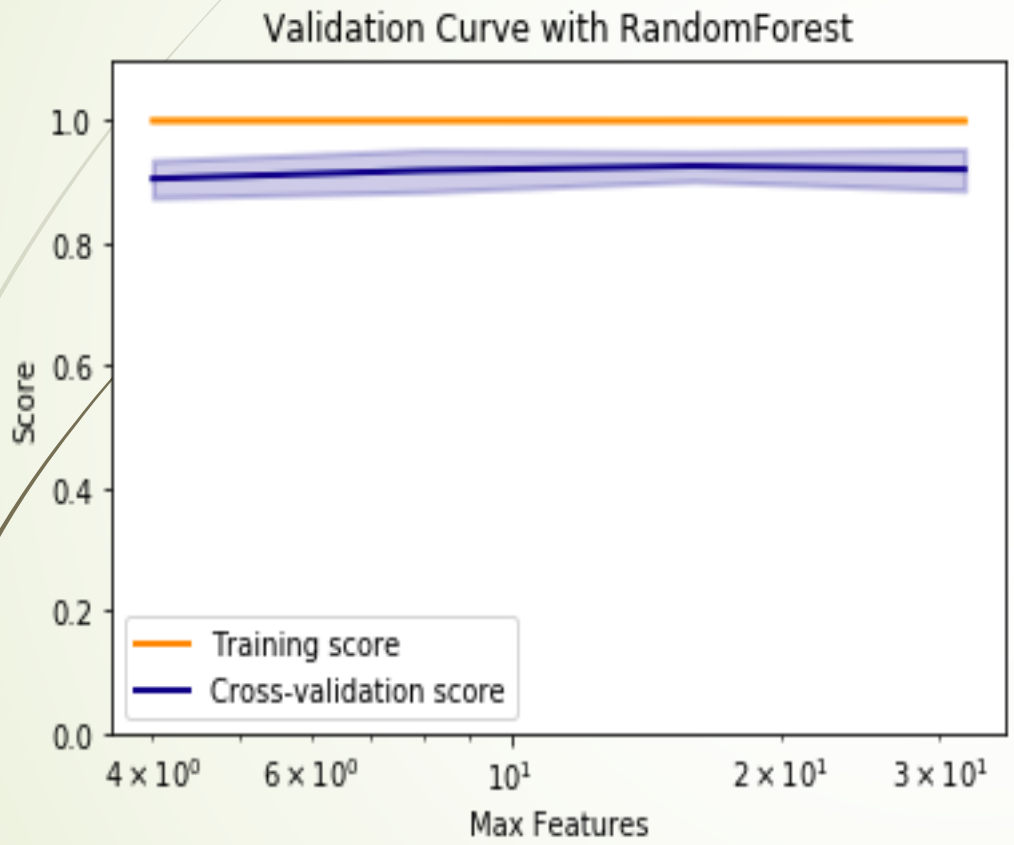
| words | importance |
|---|---|
| 2016 | 0.058426 |
| clinton | 0.045204 |
| elect | 0.036145 |
| 2017 | 0.029396 |
| hillari | 0.028819 |
| octob | 0.027306 |
| said | 0.017371 |
| cnn | 0.016577 |
| it | 0.016413 |
| politifact | 0.015954 |

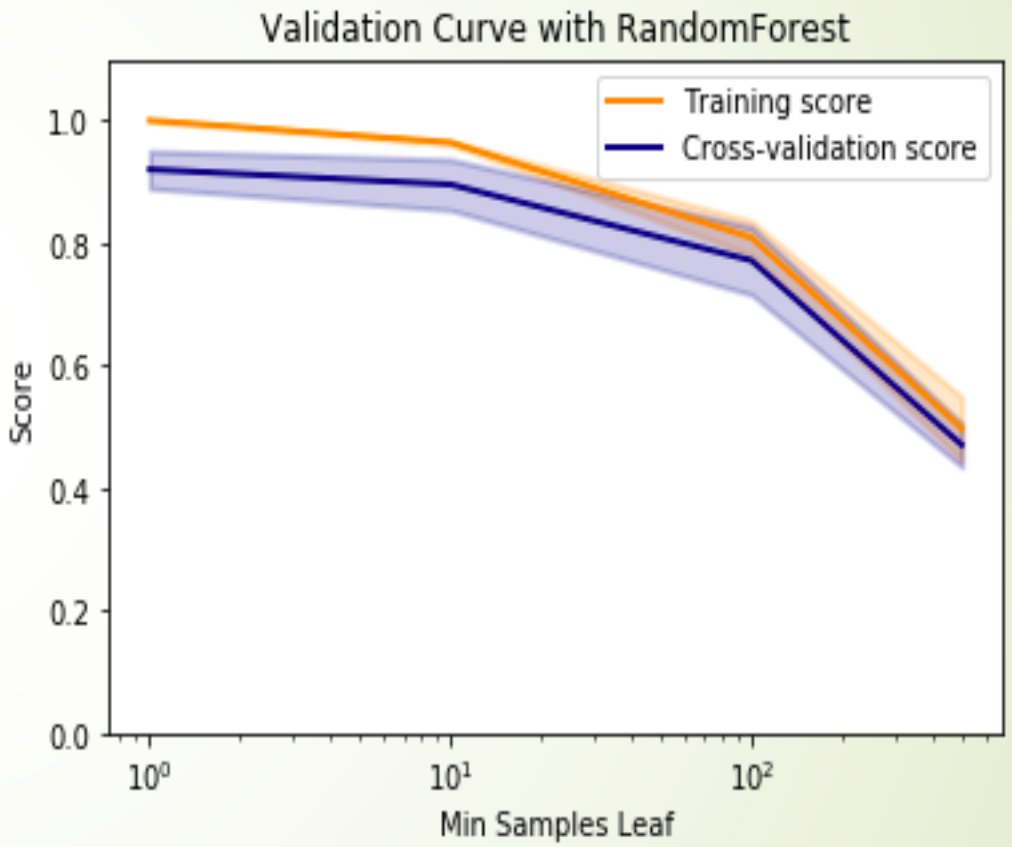
# Media Credibility Across Political Specter

| | FAKE | REAL | TOTAL | CREDIBILITY |
|---|---|---|---|---|
| **0-unknown** | 54 | 325 | 379 | 0.857520 |
| **2-left** | 105 | 1230 | 1335 | 0.921348 |
| **3-center left** | 80 | 937 | 1017 | 0.921337 |
| **4-center** | 54 | 1241 | 1295 | 0.958301 |
| **5-center right** | 5 | 126 | 131 | 0.961832 |
| **6-right** | 74 | 1360 | 1434 | 0.948396 |
| **7-extreme right** | 27 | 286 | 313 | 0.913738 |



# Validation Curve



# Conclusions:

- Using bag-of-words approach in identification of fake/real news proved as a valid approach for the given dataset and the dataset obtained through web-scraping.
- Stemming and lemmatization decreases accuracy of prediction.
- Multinomial Naive Bayes, although proven effective for predicting spam emails, did not show its effectiveness predicting fake news from real.
- RandomForest algorithm showed better results and performed well on a wide range of features.
- Importance features showed a strong influence of 2016 presidential elections

# Further Steps

- I did not exclude satire and conspiracy news from the set. Mostly because they are popular on social networks. I think that in the future those two subcategories shall be studied separately.
- It would also be interesting to re-approach this subject with the word vectorizing techniques and see what topics can be identified in respect to political affiliation of a news source.