NYU

Fake News Detector

Karanpreet Singh Wadhwa (ksw352)
Arjan Singh Narula (asn419)
Nov 26, 2019

NYU

Project Details

Bullet Points:

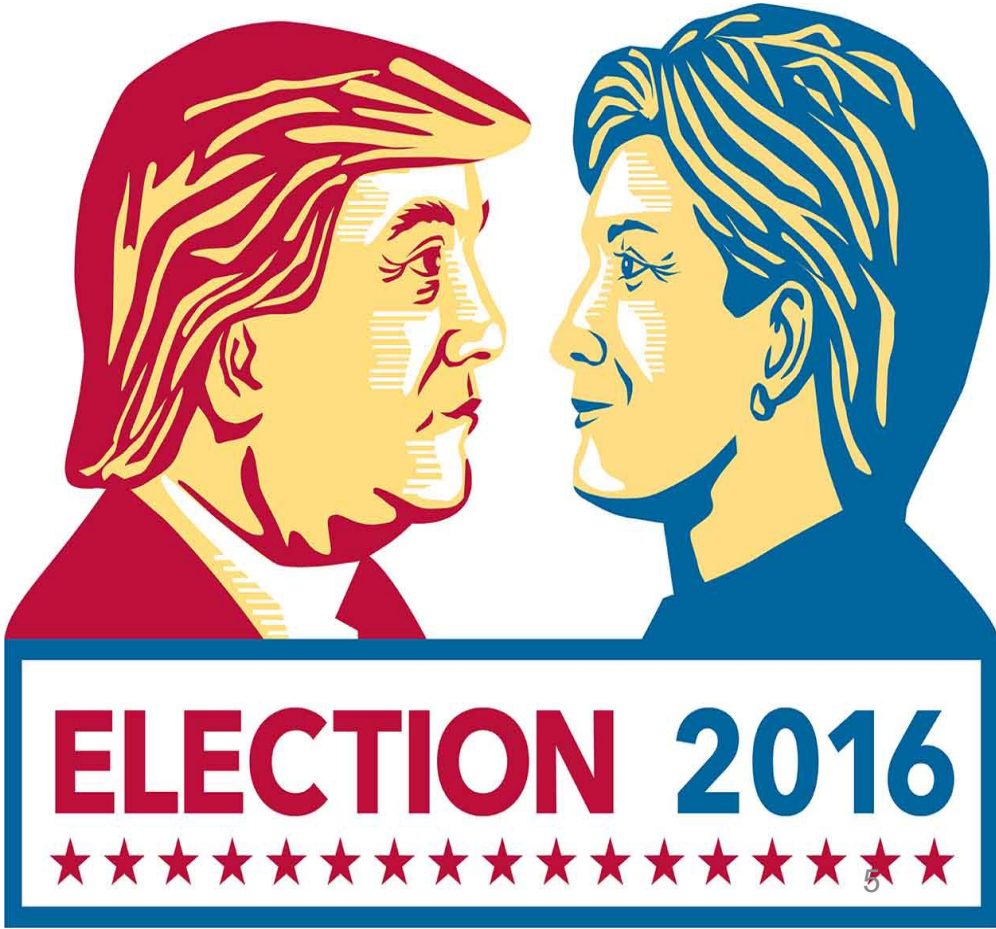
- This project is a medium with which we can deal with the dissemination of fake news.
- The **goal** of this project is to filter out fake news from the real story.
- Currently, the **Scope** of the project is to filter the news based on the title of the news.

NYU

Background

Impact:

The widespread propagation of false information online is not a recent phenomenon, but its perceived impact in the 2016 U.S. presidential election has thrust the issue into the spotlight. Apart from this, fake news has been the cause of many agitated situations and even fatalities in many countries.



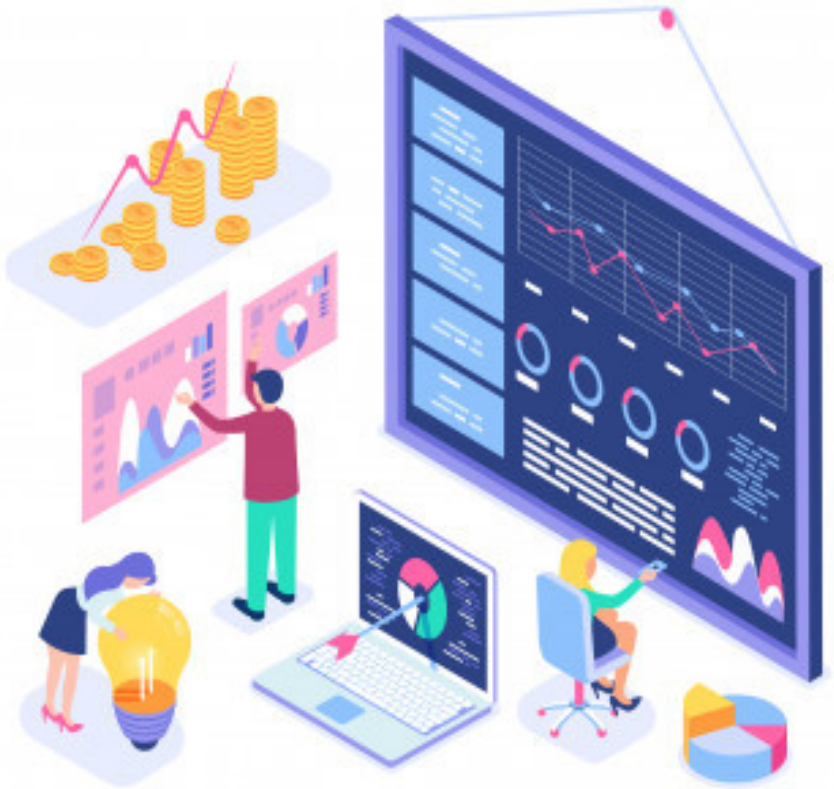
In General, The four observed flavors of “fake news”:

1) Clickbait — Shocking headlines meant to generate clicks to increase ad revenue. Frequently these stories are highly exaggerated or totally false.

2) Propaganda — Intentionally misleading or deceptive articles intended to promote the author’s agenda. Often the rhetoric is hateful and incendiary.

3) Commentary/Opinion — Biased reactions to current events. These articles frequently tell the reader how to perceive recent events.

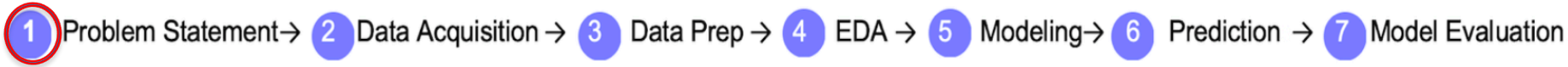
4) Humor/Satire — Articles written for entertainment. These stories are not meant to be taken seriously.



NYU

Workflow

Workflow :



Problem Statement:

- How to detect whether a given new is real or fake?
- Are there any words or groups of words (Probably Bait) that are plentiful in fake news?

Workflow :

Data Acquisition:

- Web Scrapping news data from Reddit and Polifact.

	Unnamed: 0	author	domain	num_comments	score	subreddit	timestamp	title
0	0	Redditissold	politics.theonion.com	17	1	TheOnion	1574614835	Trump honors war criminal with presidential string of human ears
1	1	antdude	sports.theonion.com	0	1	TheOnion	1574580183	Kyrie Irving Debuts Signature Shoe Inspired By RFID Chips Government Secretly Implants In Anesthetized Patients
2	2	Sanlear	local.theonion.com	11	1	TheOnion	1574519785	Veterinarian Wishes Owner Would Just Let Dog Answer One Goddamn Question
3	3	EpicBroomGuy	politics.theonion.com	0	1	TheOnion	1574482602	'I Could Spare Some Change,' Says Man About To Become Buttigieg Campaign's Top Black Donor
4	4	PeopleNeedPower	entertainment.theonion.com	5	1	TheOnion	1574471849	Smiling, Knife-Wielding Marie Kondo Orders Followers To Leave Behind Cluttered Physical Forms

Workflow :

Data Preparation: (Major)

- Remove punctation
- Remove numbers
- Transform all text to lowercase
- Remove English Stop words.

Before

title
Trump honors war criminal with presidential string of human ears
Kyrie Irving Debuts Signature Shoe Inspired By RFID Chips Government Secretly Implants In Anesthetized Patients
Veterinarian Wishes Owner Would Just Let Dog Answer One Goddamn Question
'I Could Spare Some Change,' Says Man About To Become Buttigieg Campaign's Top Black Donor
Smiling, Knife-Wielding Marie Kondo Orders Followers To Leave Behind Cluttered Physical Forms

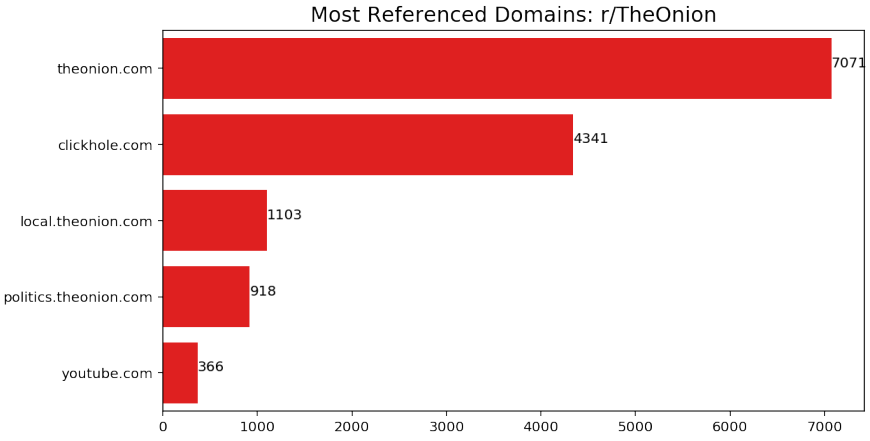
After

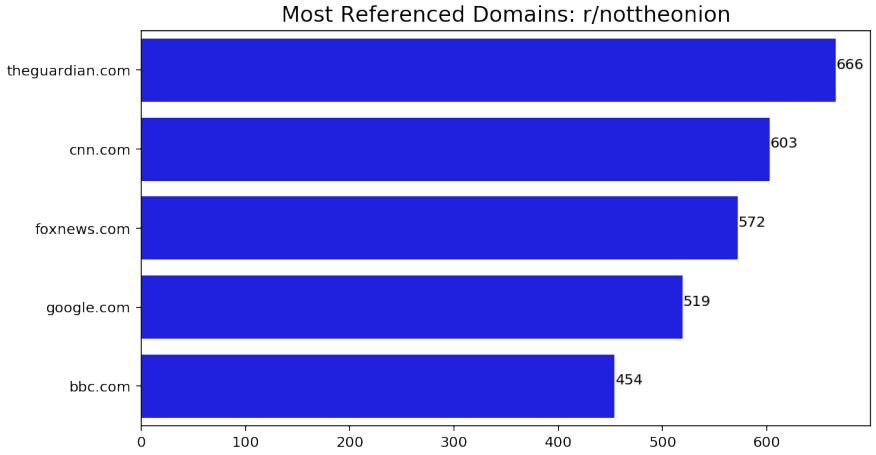
title
trump honors war criminal with presidential string of human ears
kyrie irving debuts signature shoe inspired by rfid chips government secretly implants in anesthetized patients
veterinarian wishes owner would just let dog answer one goddamn question
i could spare some change says man about to become buttigieg campaign s top black donor
smiling knife wielding marie kondo orders followers to leave behind cluttered physical forms

Workflow :

1 Problem Statement → 2 Data Acquisition → 3 Data Prep → 4 EDA → 5 Modeling → 6 Prediction → 7 Model Evaluation

EDA of Redditt Data based on **Domain** :

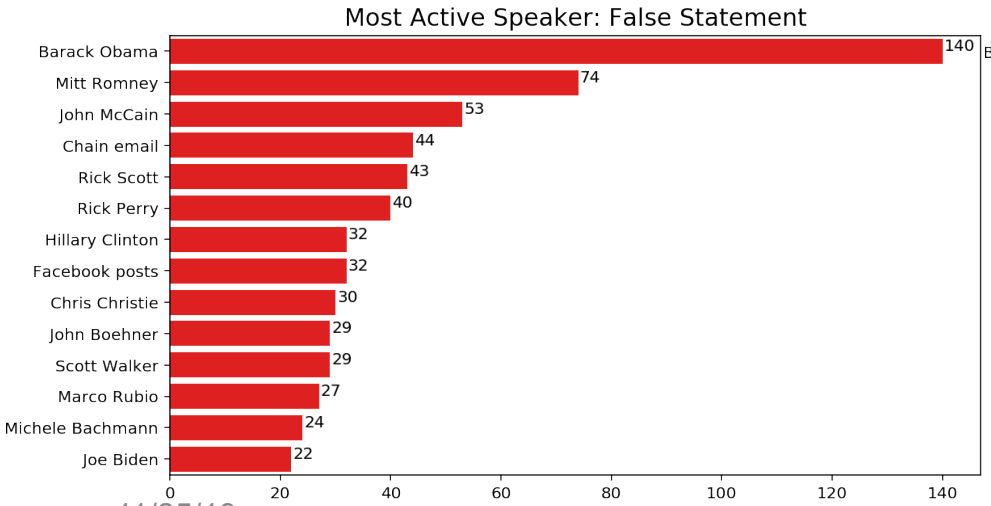


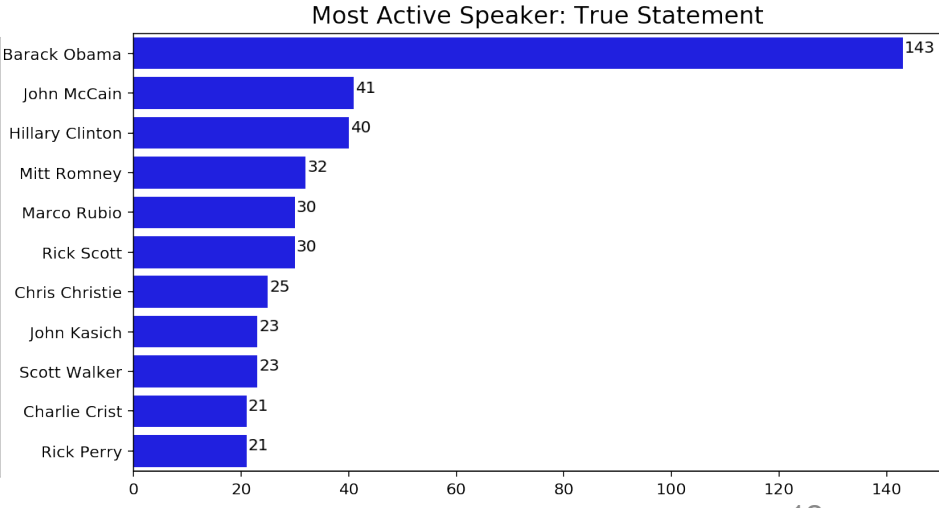


Workflow :

1 Problem Statement → 2 Data Acquisition → 3 Data Prep → 4 EDA → 5 Modeling → 6 Prediction → 7 Model Evaluation

EDA of Polifact Data based on **Speaker** :

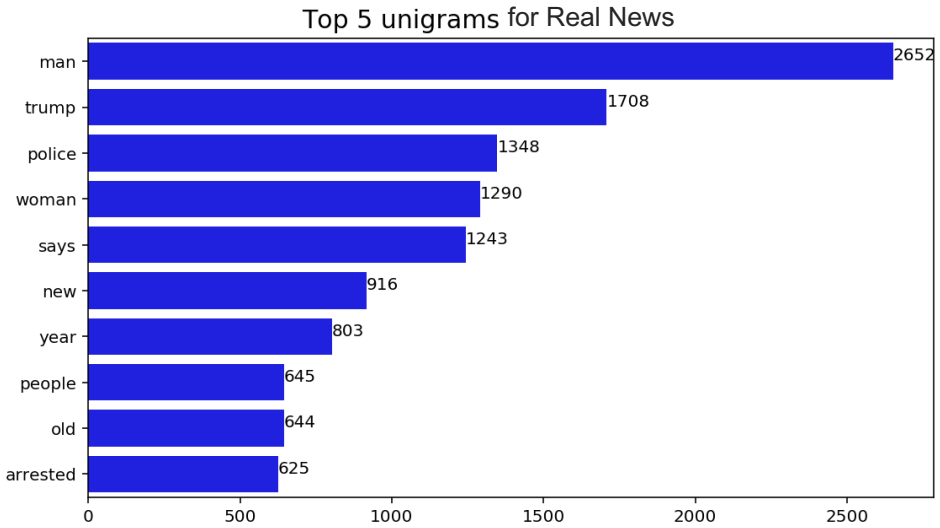


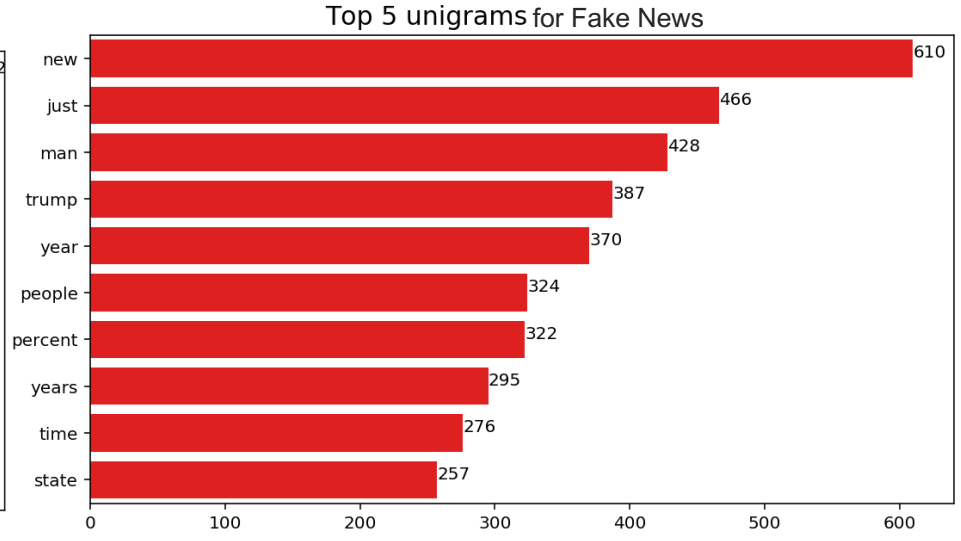


Workflow :

1 Problem Statement → 2 Data Acquisition → 3 Data Prep → 4 EDA → 5 Modeling → 6 Prediction → 7 Model Evaluation

EDA after CountVectorizer with n_grams (1,1) :

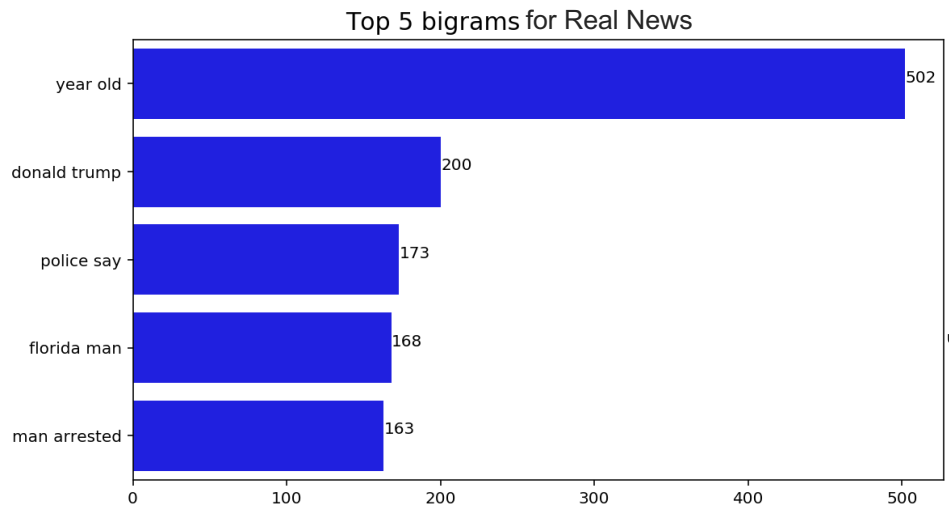


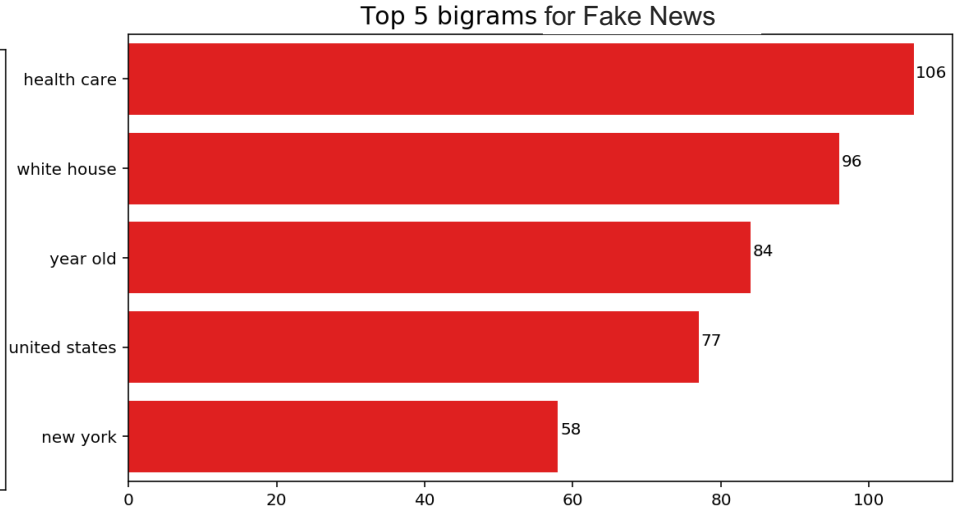


Workflow :

1 Problem Statement → 2 Data Acquisition → 3 Data Prep → 4 EDA → 5 Modeling → 6 Prediction → 7 Model Evaluation

EDA after CountVectorizer with n_grams (2,2) :

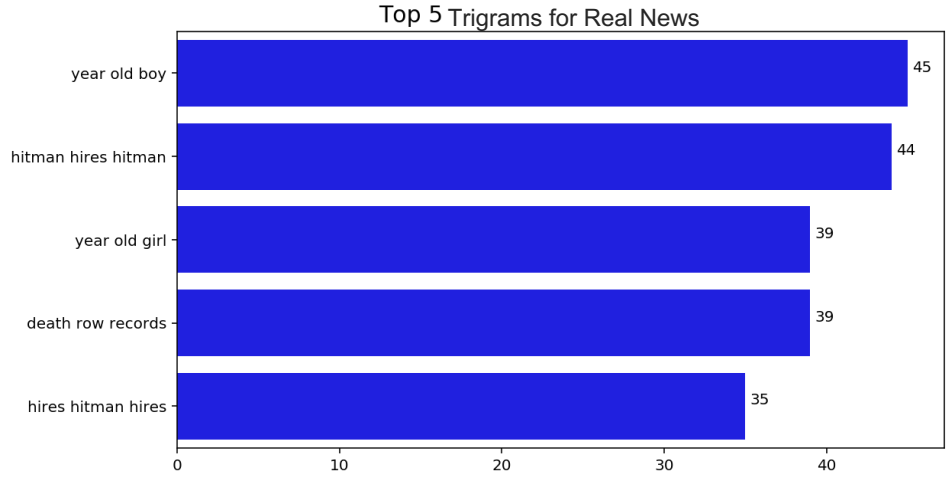


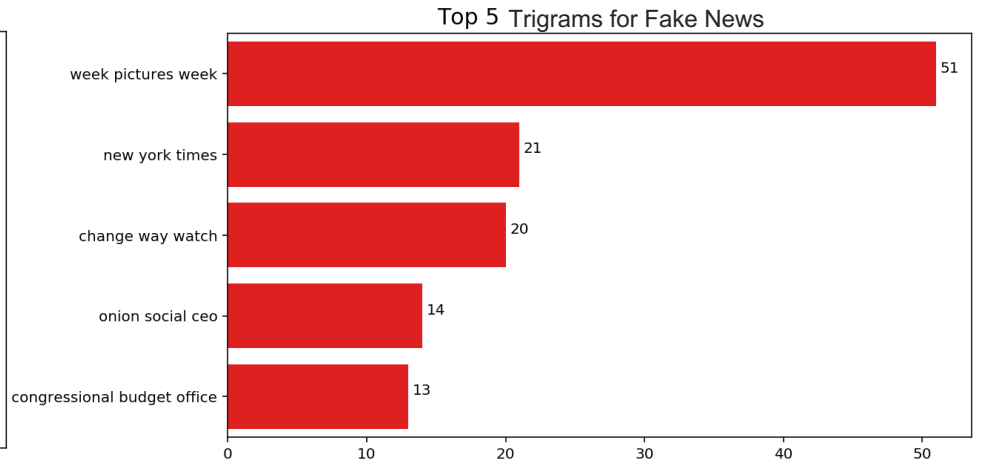


Workflow :

1 Problem Statement → 2 Data Acquisition → 3 Data Prep → 4 EDA → 5 Modeling → 6 Prediction → 7 Model Evaluation

EDA after CountVectorizer with n_grams (3,3) :





1 Problem Statement → 2 Data Acquisition → 3 Data Prep → 4 EDA → 5 Modeling → 6 Prediction → 7 Model Evaluation

Model 1 (CountVectorizer & Logistic Regression) :

```
pipe = Pipeline([('cvec', CountVectorizer()),
                 ('lr', LogisticRegression(solver='liblinear'))])

# Different Parameters to try
pipe_params = {'cvec__stop_words': [None, 'english'],
               'cvec__ngram_range': [(1,1), (2,2), (1,3)],
               'lr__C': [0.01, 1]}
# GridSearch to find best
gs = GridSearchCV(pipe, param_grid=pipe_params, cv=3)
gs.fit(X_train, y_train);
print("Best score:", gs.best_score_)
print("Train score", gs.score(X_train, y_train))
print("Test score", gs.score(X_test, y_test))
print("Best paramenters are :{}".format(gs.best_params_))

```

executed in 48.5s, finished 22:14:13 2019-11-25

```
Best score: 0.8199780795226207
Train score 0.9347256895816842
Test score 0.8262855055256187
Best paramenters are :{'cvec__ngram_range': (1, 1), 'cvec__stop_words': None, 'lr__C': 1}
```

Workflow :

1 Problem Statement → 2 Data Acquisition → 3 Data Prep → 4 EDA → 5 Modeling → 6 Prediction → 7 Model Evaluation

Model 1 Evaluation:

Accuracy: 82.63 %
Precision: 78.44 %
Recall: 81.09 %
Specificity: 83.75 %
Misclassification Rate: 15.2 %

VS

Base Accuracy: (Percent of class in the data)

Real News 0.58%
Fake News 0.42%

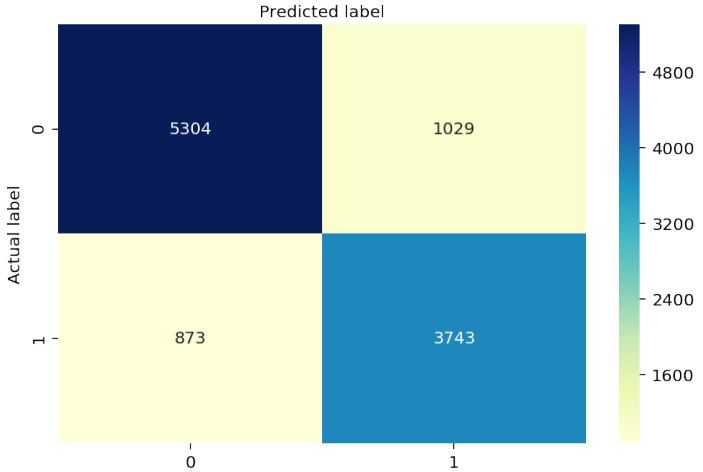
Workflow :

1 Problem Statement → 2 Data Acquisition → 3 Data Prep → 4 EDA → 5 Modeling → 6 Prediction → 7 Model Evaluation

Model 1 Evaluation:

Accuracy: 82.63 %
Precision: 78.44 %
Recall: 81.09 %
Specificity: 83.75 %
Misclassification Rate: 15.2 %

Confusion Matrix:



Real News : 0, Fake News : 1

Milestone Completed:

- Extracted Satire and political news Data.
- Completed whole Workflow cycle for one model.

Milestone to complete:

- Extract data from Propaganda News for (Activist Report, Natural News), Hoax News from (DC Gazette, American News) and satire news from (Clickhole).
- Implement TfidfVectorizer for Logistic regression.
- Implement CountVextorizer and TfidfVectorizer with Naïve Bayes and Multinomial Naïve Bayes.
- Choose final model.