Fake News Detector

Karanpreet Singh Wadhwa (ksw352)
Arjan Singh Narula (asn419)
DEC 09 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.

- **Class Scope:** We have Studied text analysis in Lecture 8 and 9 of the Class, and we would be utilizing those concepts.

*

NYU

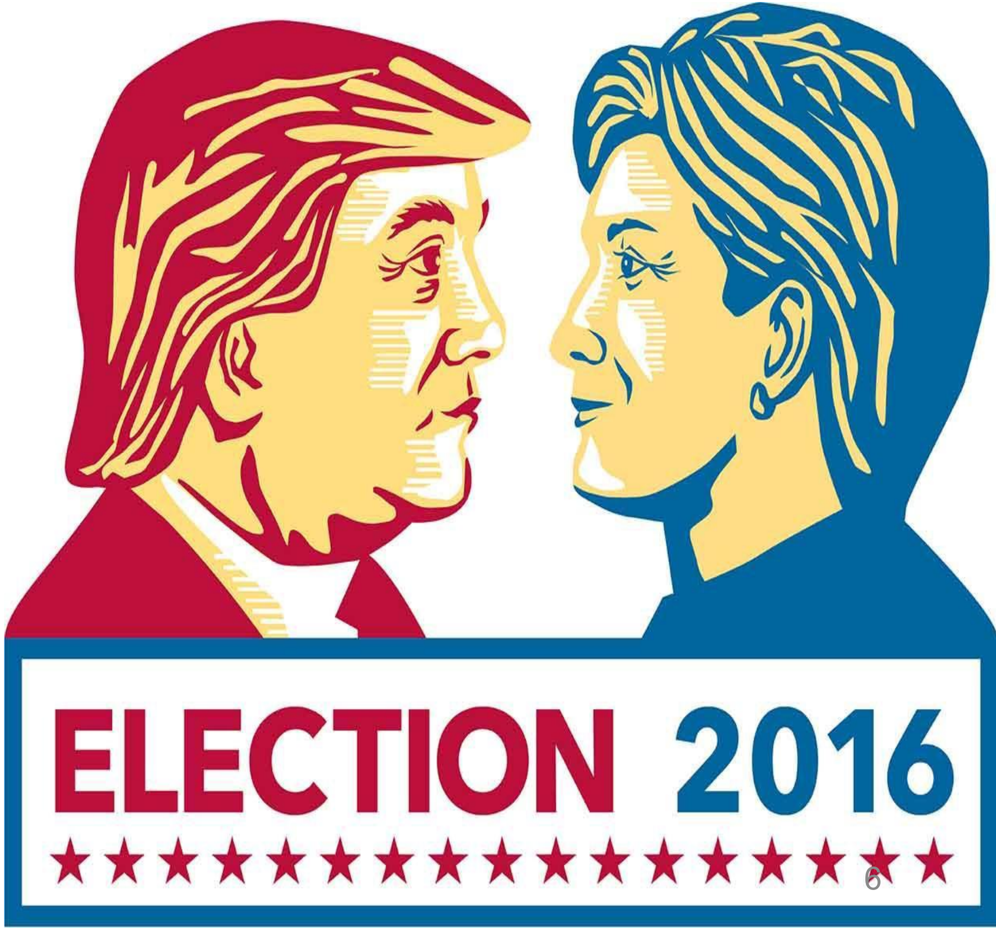
Background

Background:

Fake news is now viewed as one of the greatest threats to democracy, journalism, and freedom of expression. It has weakened public trust in governments and its potential impact on the contentious “Brexit” referendum and the equally divisive 2016 U.S. presidential election – which it might have affected– is yet to be realized. The reach of fake news was best highlighted during the critical months of the 2016 U.S. presidential election campaign, where the top twenty frequently-discussed false election stories generated 8,711,000 shares, reactions, and comments on Facebook, ironically, larger than the total of 7,367,000 for the top twenty most-discussed election stories posted by 19 major news websites. Our economies are not immune to the spread of fake news either, with fake news being connected to stock market fluctuations and massive trades. For example, fake news claiming that Barack Obama was injured in an explosion wiped out $130 billion in stock value. These events and losses have motivated fake news research and sparked the discussion around fake news, as observed by skyrocketing usage of terms such as “post-truth” – selected as the international word of the year by Oxford Dictionaries in 2016.

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.



Types of Fake News

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.

WHAT IS FAKE NEWS?

And how can you spot it?

Fake news is a deliberately made up story which aims to get people to believe something that is not true, or a story that may mislead you because it is not completely accurate.

1 *Find out who published it. Real news is published by trustworthy news sources with a reputation for checking their facts, like the BBC.*

2 *What claims does the news story make? Are these backed up by facts or evidence that you can double check? If you can't find any facts or evidence, take care – this might be a fake news story.*

3 *Does the story include more than one point of view? Fake news stories often only present one side of the story.*

4 *How does the story make you feel? If the story makes you feel angry or frightened, be careful. Fake news stories often use your emotions to try to influence what you believe.*

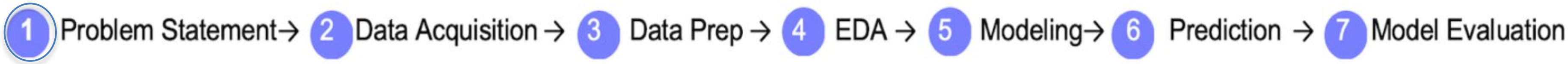
5 *If you think a story might be fake news, then do more more research. Ask more questions and find answers from another source.*

gresham-books.co.uk

NYU

Workflow

Workflow :



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

Problem Statement:

- 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?

NYU

Data Acquisition

Workflow :

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

Data Acquisition:

- Sites Data is Extracted from.
  1) Reddit
  2) The Guardian
  3) Daily Mail
- Sites Dataset Gathered/picked from.
  1. Gossipcop
  2. Politifact
  3. Washington University Fake News dataset

Workflow :

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

Data Acquisition:

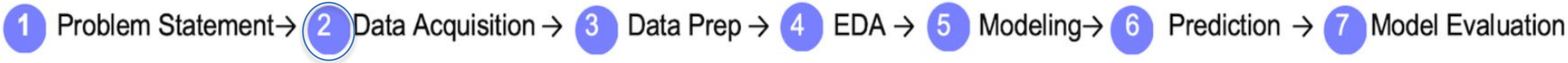
- Fake News:

1. TheOnion (Reddit) Contains data about absurd, satirical Fake news.
2. Politifact contains data about propaganda and hoax news.
3. Gossipcop contains data about Entertainment Fake news.
4. Washington University Fake News Dataset consists of all types of Fake news.

NYU

Data Acquisition

Workflow :

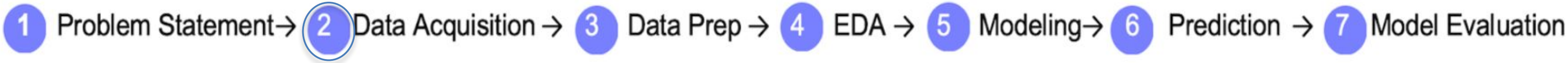

1 Problem Statement→ 2 Data Acquisition → 3 Data Prep → 4 EDA → 5 Modeling→ 6 Prediction → 7 Model Evaluation

Data Acquisition:

- Real News:

1. NotTheOnion contains data about absurd, Satirical Real news.
2. TheGuardian contains data about all sector of real news from politics to sports.
3. Daily Mail contains data about all sector of real news from politics to sports.
4. Washington University contains data about all sector of real news from politics to sports.

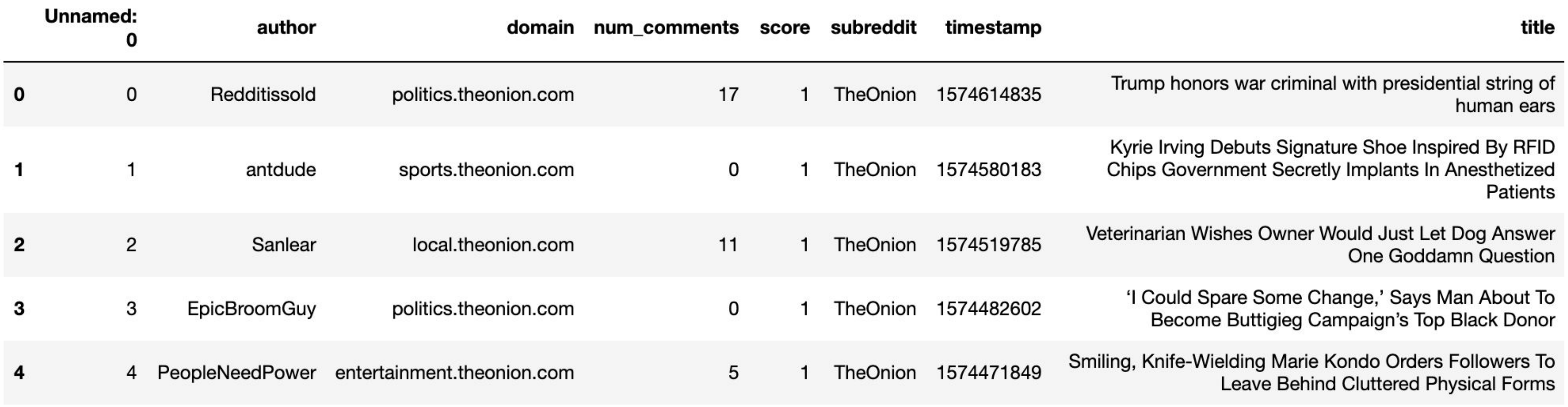
Workflow :



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

Data Acquisition:

- Web Scraping news data from Reddit.



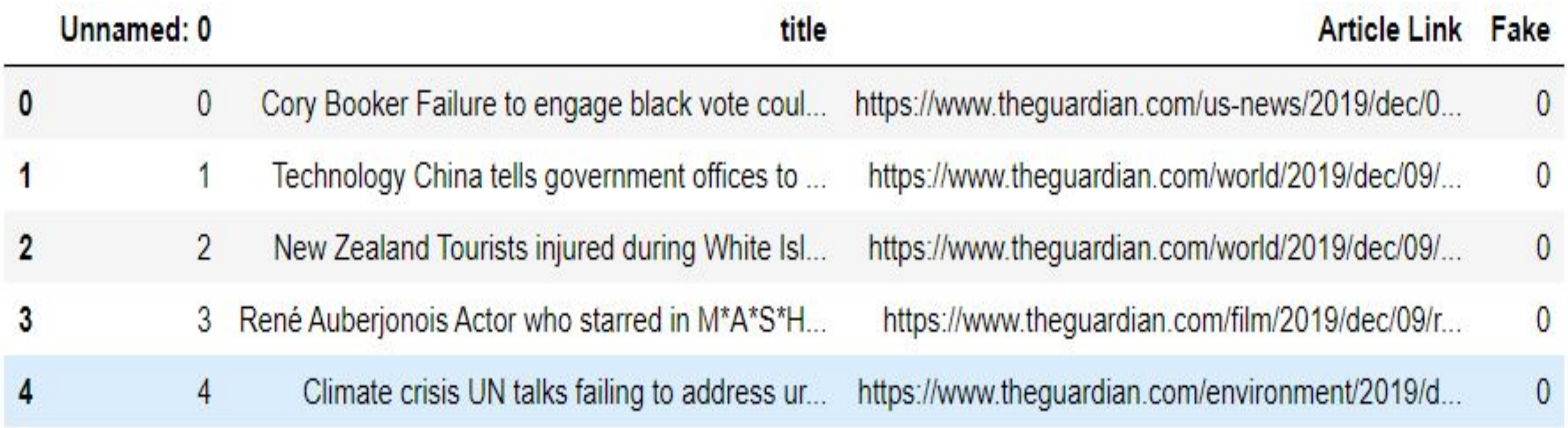
	Unnamed: 0	author	domain	num_comments	score	subreddit	timestamp	title
0	0	Redditissold	politics.theonion.com	17	1	TheOnion	1574614835	Trump honors war criminal with presidential string of human ears
1	1	antdude	sports.theonion.com	0	1	TheOnion	1574580183	Kyrie Irving Debuts Signature Shoe Inspired By 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 :

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

Data Acquisition:

- Web Scraping news data from The Guardian.



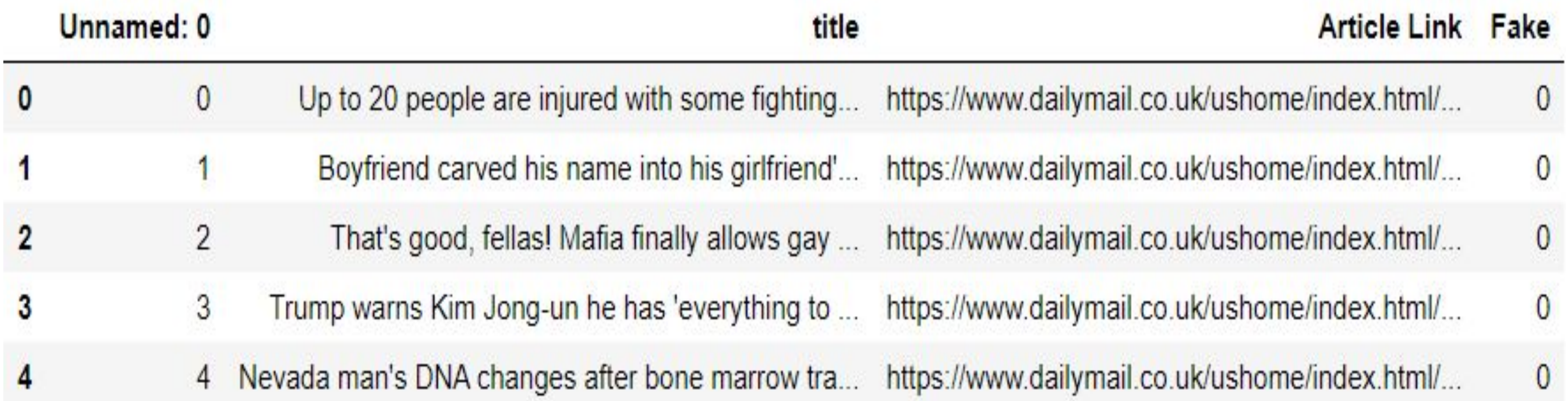
	Unnamed: 0	title	Article Link	Fake
0	0	Cory Booker Failure to engage black vote coul...	https://www.theguardian.com/us-news/2019/dec/0...	0
1	1	Technology China tells government offices to ...	https://www.theguardian.com/world/2019/dec/09/...	0
2	2	New Zealand Tourists injured during White Isl...	https://www.theguardian.com/world/2019/dec/09/...	0
3	3	René Auberjonois Actor who starred in M*A*S*H...	https://www.theguardian.com/film/2019/dec/09/r...	0
4	4	Climate crisis UN talks failing to address ur...	https://www.theguardian.com/environment/2019/d...	0

Workflow :

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

Data Acquisition:

- Web Scraping news data from Daily Mail.



	Unnamed: 0	title	Article Link	Fake
0	0	Up to 20 people are injured with some fighting...	https://www.dailymail.co.uk/ushome/index.html/...	0
1	1	Boyfriend carved his name into his girlfriend'...	https://www.dailymail.co.uk/ushome/index.html/...	0
2	2	That's good, fellas! Mafia finally allows gay ...	https://www.dailymail.co.uk/ushome/index.html/...	0
3	3	Trump warns Kim Jong-un he has 'everything to ...	https://www.dailymail.co.uk/ushome/index.html/...	0
4	4	Nevada man's DNA changes after bone marrow tra...	https://www.dailymail.co.uk/ushome/index.html/...	0

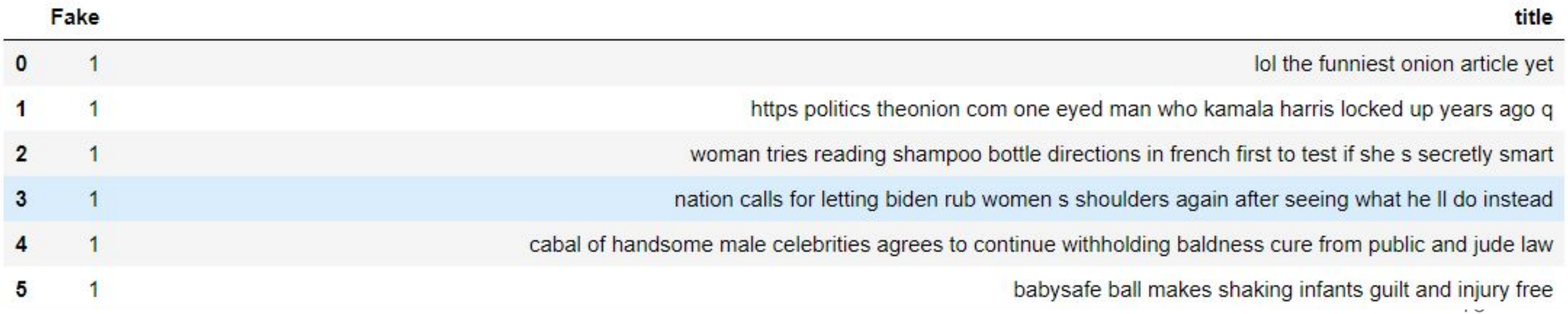
Data Acquisition

Workflow :

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

Data Acquisition:

- **Target Variable** :
  - “Fake” - Binary variable (1 for fake news and 0 for real news).
- **Predictor Variable** :
  - “title”- Text (contains English text.)



	Fake	title
0	1	lol the funniest onion article yet
1	1	https politics theonion com one eyed man who kamala harris locked up years ago q
2	1	woman tries reading shampoo bottle directions in french first to test if she s secretly smart
3	1	nation calls for letting biden rub women s shoulders again after seeing what he ll do instead
4	1	cabal of handsome male celebrities agrees to continue withholding baldness cure from public and jude law
5	1	babysafe ball makes shaking infants guilt and injury free

Data Preparation

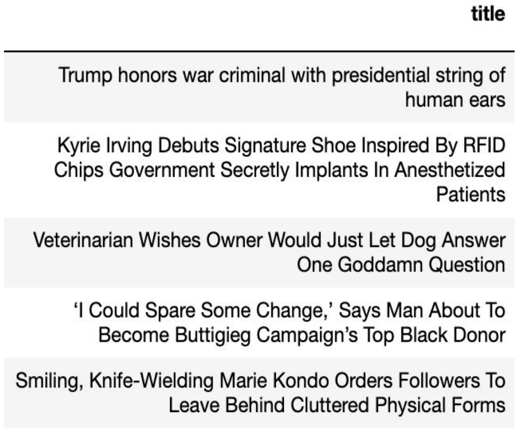
Workflow :

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

Data Preparation: (Major)

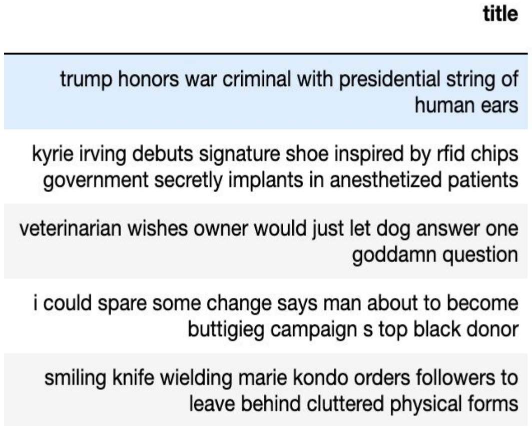
- Remove punctuation
- Remove numbers
- Transform all text to lowercase
- Remove English Stop words.

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

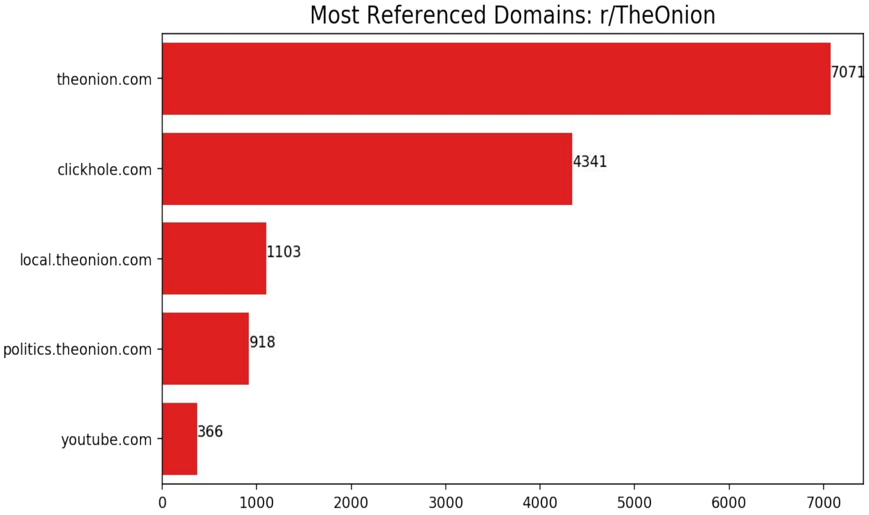
Exploratory Data Analysis:

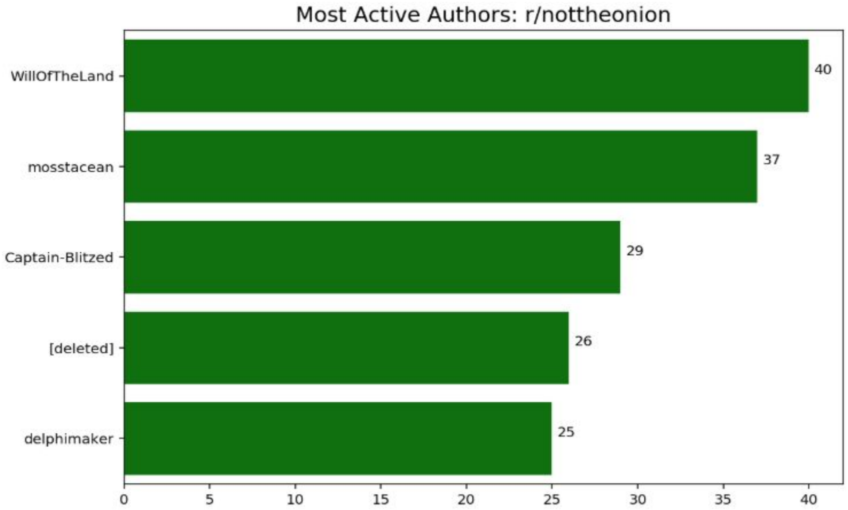
Exploratory data analysis (EDA) is a term for certain kinds of initial analysis and findings done with data sets, usually early on in an analytical process. Some experts describe it as “taking a peek” at the data to understand more about what it represents and how to apply it. Exploratory data analysis is often a precursor to other kinds of work with statistics and data.

Workflow :

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

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



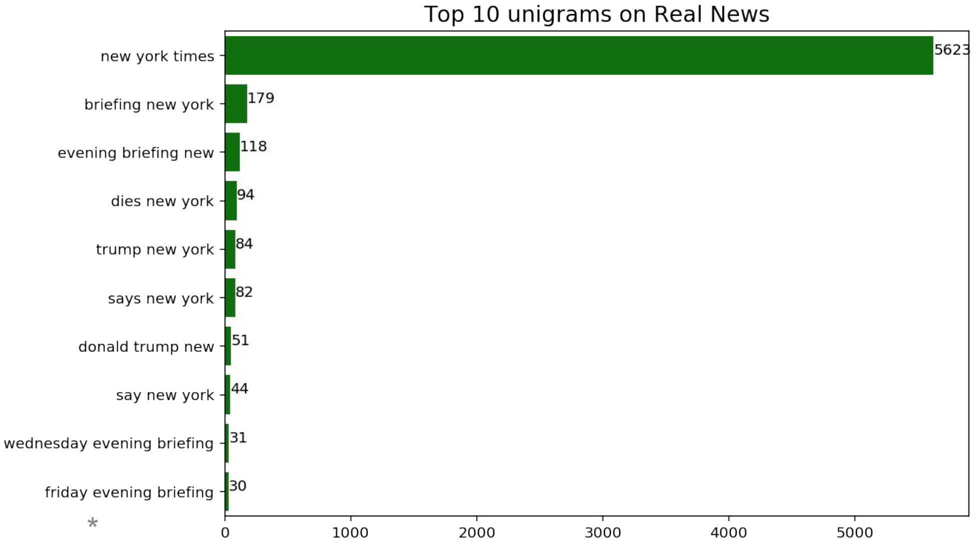


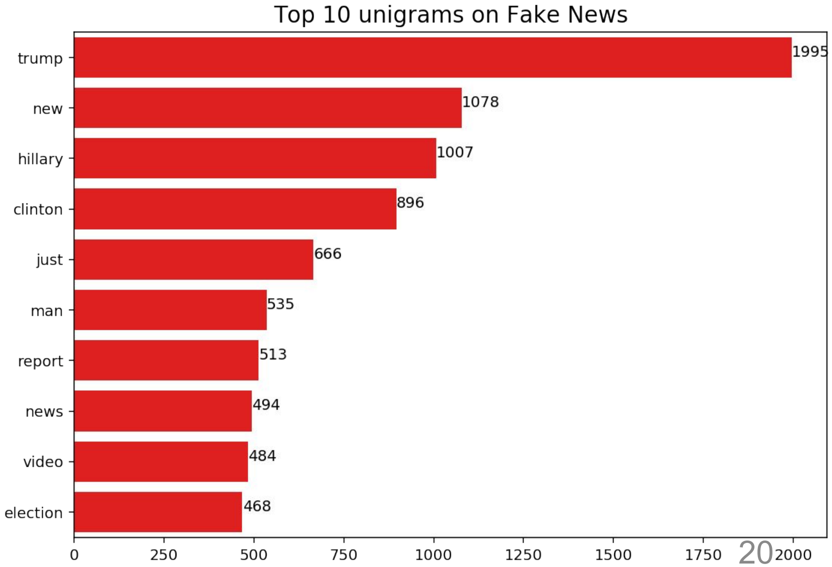
*

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) :





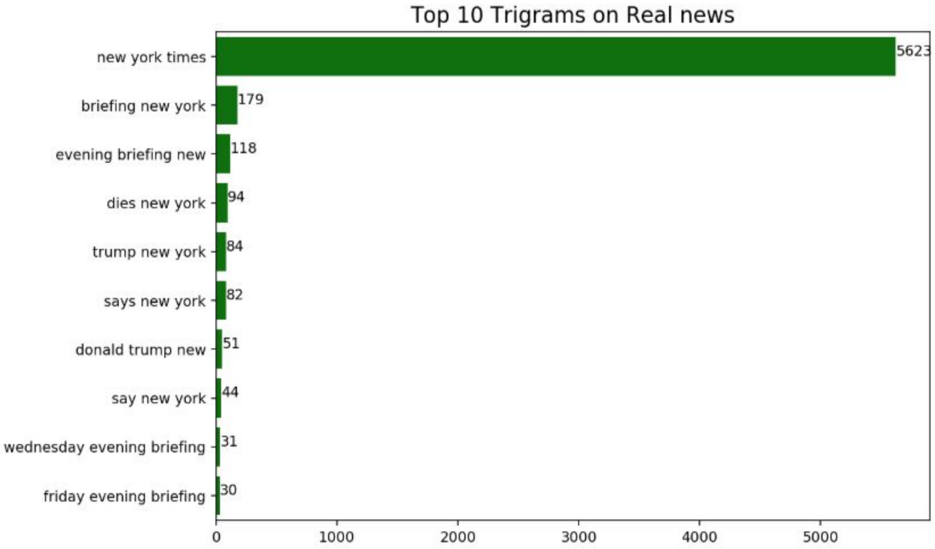
*

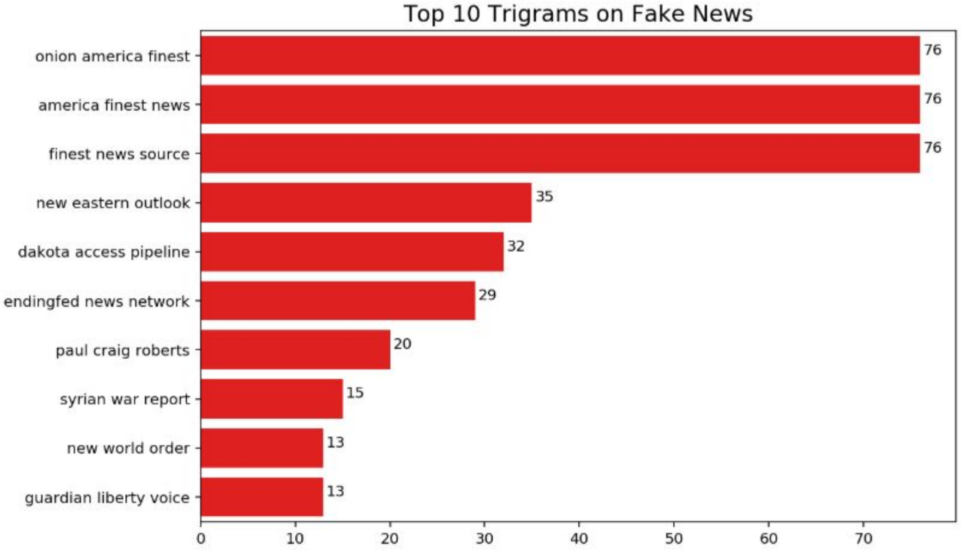
EDA

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) :





Modeling & Prediction

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

NLP Techniques:(To transform text in suitable form for the model)

1. CountVectorizer
2. TfIdfVectorizer

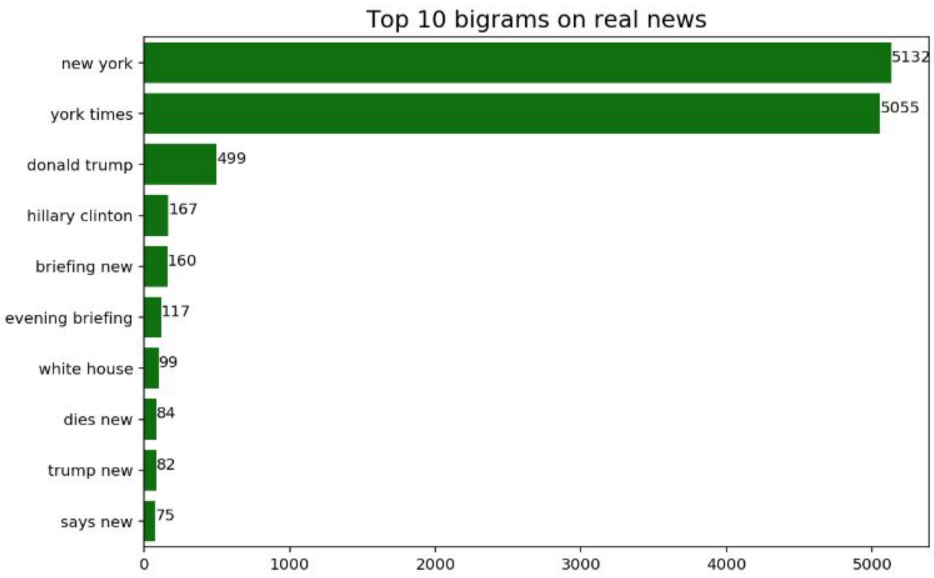
Model:(These Models work best with the text. So we would use these models)

1. Logistic Regression
2. Multinomial Naive Bayes

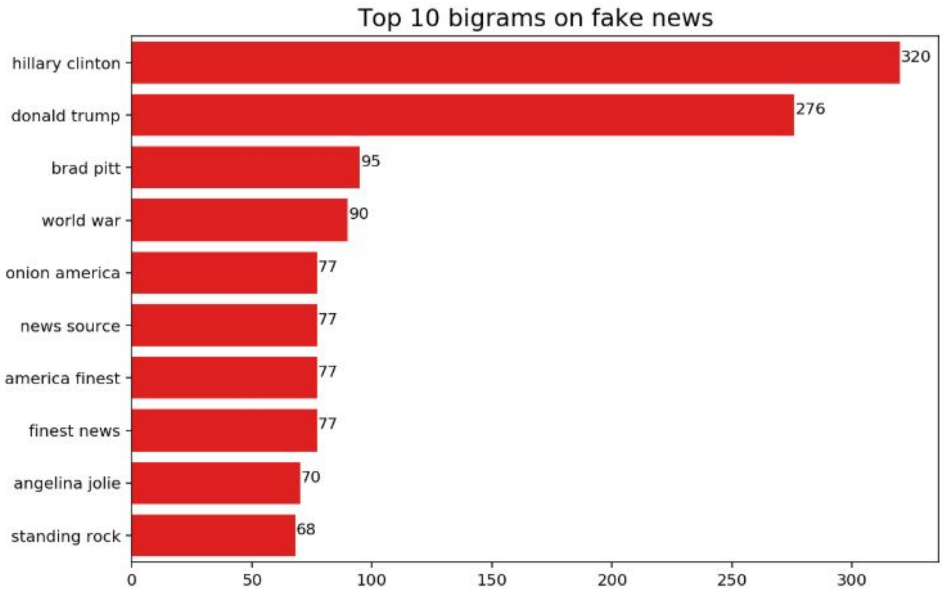
Workflow :

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) :



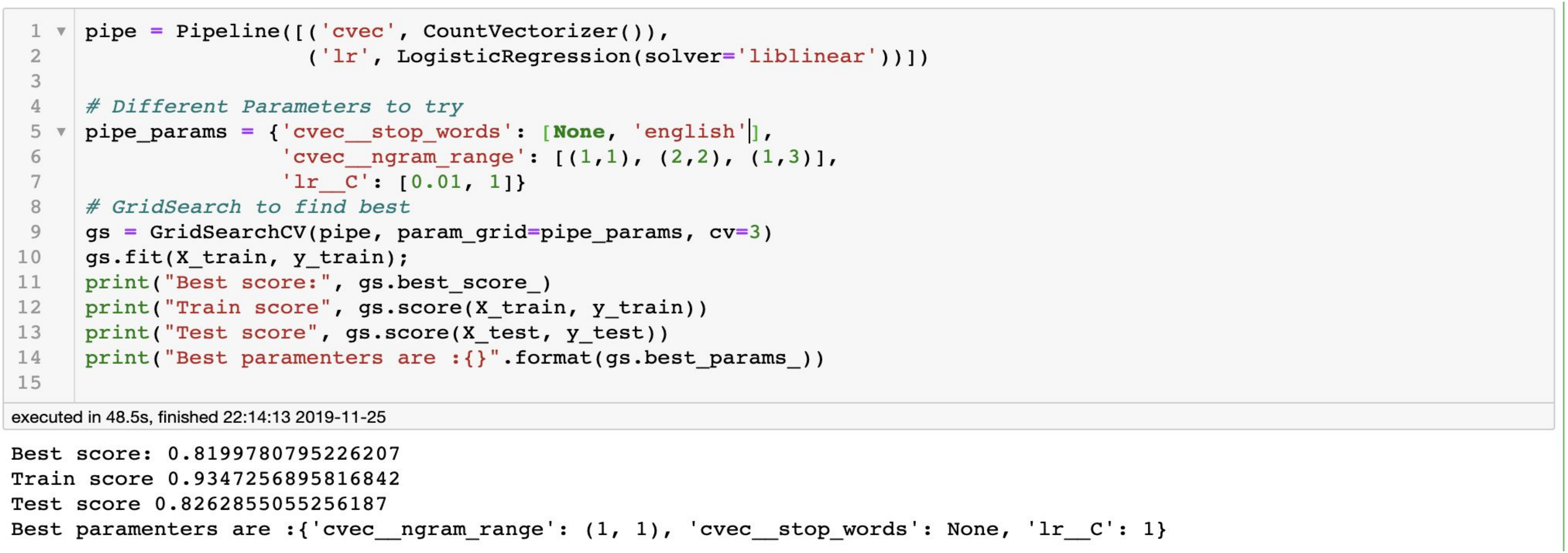
Top 10 bigrams on real news



Top 10 bigrams on fake news

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

Model 1 (Logistic Regression and CountVectorizer) :



```
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}
```

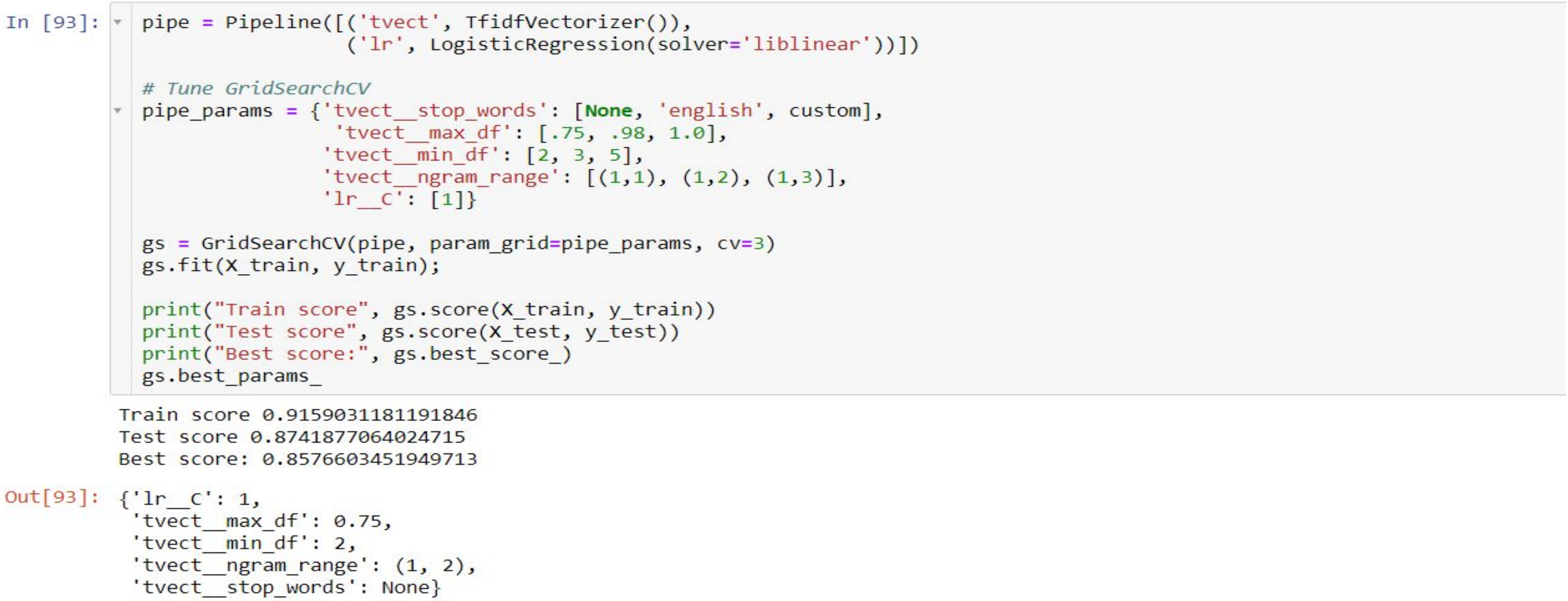
*

NYU

Modeling & Prediction

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

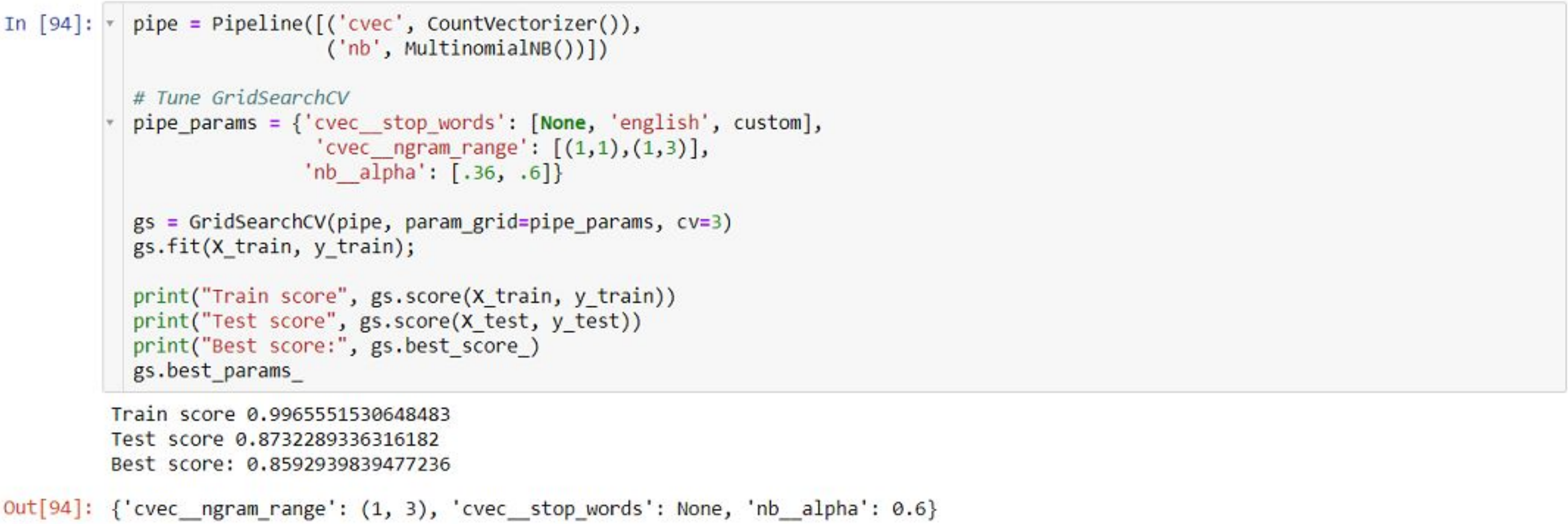
Model 2 (Logistic Regression and TfidfVectorize) :



```
In [93]: pipe = Pipeline([('tvect', TfidfVectorizer()),
                          ('lr', LogisticRegression(solver='liblinear'))])

         # Tune GridSearchCV
         pipe_params = {'tvect__stop_words': [None, 'english', custom],
                         'tvect__max_df': [.75, .98, 1.0],
                        'tvect__min_df': [2, 3, 5],
                        'tvect__ngram_range': [(1,1), (1,2), (1,3)],
                        'lr__C': [1]}

         gs = GridSearchCV(pipe, param_grid=pipe_params, cv=3)
         gs.fit(X_train, y_train);

         print("Train score", gs.score(X_train, y_train))
         print("Test score", gs.score(X_test, y_test))
         print("Best score:", gs.best_score_)
         gs.best_params_

        Train score 0.9159031181191846
        Test score 0.8741877064024715
        Best score: 0.8576603451949713

Out[93]: {'lr__C': 1,
          'tvect__max_df': 0.75,
          'tvect__min_df': 2,
          'tvect__ngram_range': (1, 2),
          'tvect__stop_words': None}
```

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

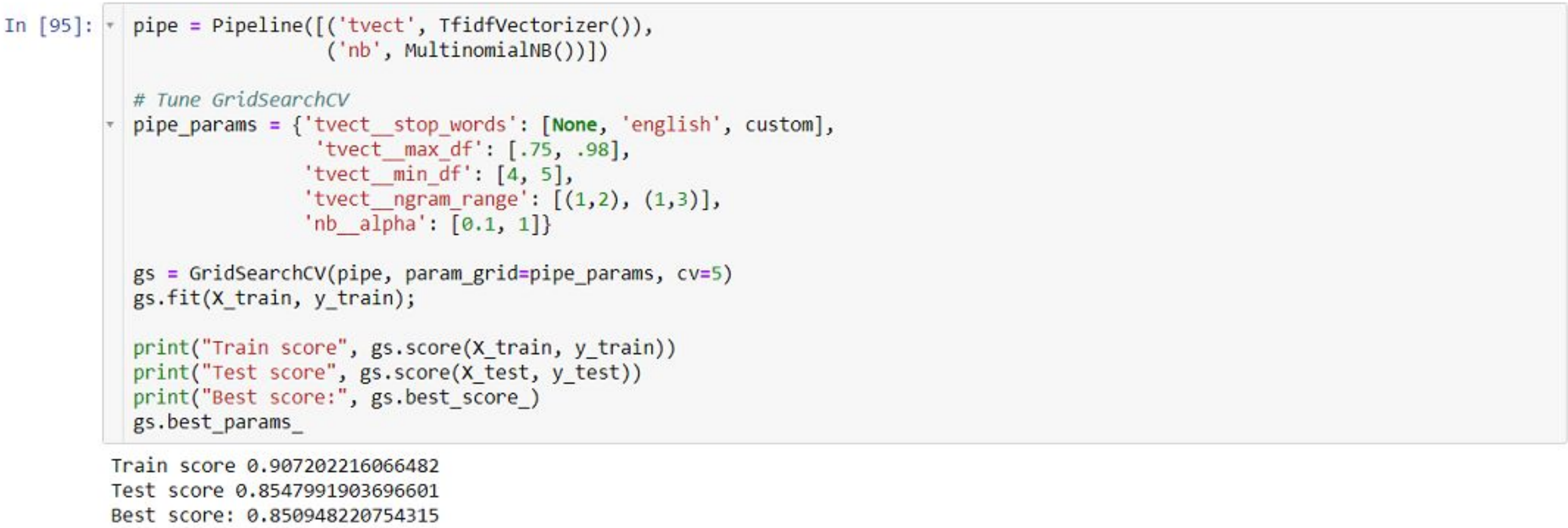
Model 3 (MultinomialNB and CountVectorizer) :



```
In [94]: pipe = Pipeline([('cvec', CountVectorizer()),
                          ('nb', MultinomialNB())])

         # Tune GridSearchCV
         pipe_params = {'cvec__stop_words': [None, 'english', custom],
                         'cvec__ngram_range': [(1,1),(1,3)],
                        'nb__alpha': [.36, .6]}

         gs = GridSearchCV(pipe, param_grid=pipe_params, cv=3)
         gs.fit(X_train, y_train);

         print("Train score", gs.score(X_train, y_train))
         print("Test score", gs.score(X_test, y_test))
         print("Best score:", gs.best_score_)
         gs.best_params_

        Train score 0.9965551530648483
        Test score 0.8732289336316182
        Best score: 0.8592939839477236

Out[94]: {'cvec__ngram_range': (1, 3), 'cvec__stop_words': None, 'nb__alpha': 0.6}
```

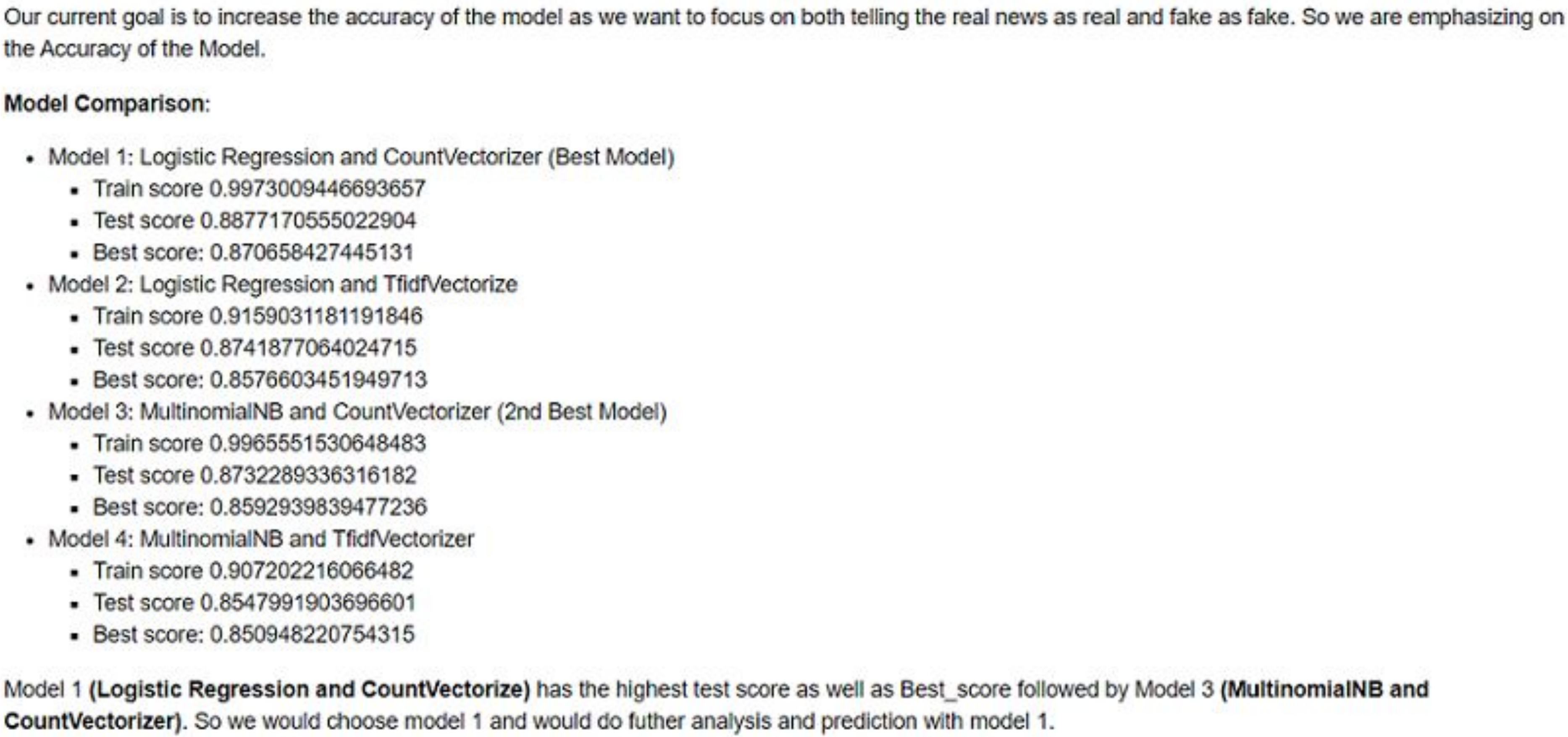
1 Problem Statement→ 2 Data Acquisition → 3 Data Prep → 4 EDA → 5 Modeling→ 6 Prediction → 7 Model Evaluation

Model 4 (CountVectorizer & Logistic Regression) :



```
In [95]: pipe = Pipeline([('tvect', TfidfVectorizer()),
                          ('nb', MultinomialNB())])

         # Tune GridSearchCV
         pipe_params = {'tvect__stop_words': [None, 'english', custom],
                         'tvect__max_df': [.75, .98],
                        'tvect__min_df': [4, 5],
                        'tvect__ngram_range': [(1,2), (1,3)],
                        'nb__alpha': [0.1, 1]}

         gs = GridSearchCV(pipe, param_grid=pipe_params, cv=5)
         gs.fit(X_train, y_train);

         print("Train score", gs.score(X_train, y_train))
         print("Test score", gs.score(X_test, y_test))
         print("Best score:", gs.best_score_)
         gs.best_params_

Train score 0.907202216066482
Test score 0.8547991903696601
Best score: 0.850948220754315
```

Model Evaluation



Our current goal is to increase the accuracy of the model as we want to focus on both telling the real news as real and fake as fake. So we are emphasizing on the Accuracy of the Model.

Model Comparison:

- Model 1: Logistic Regression and CountVectorizer (Best Model)
  - Train score 0.9973009446693657
  - Test score 0.8877170555022904
  - Best score: 0.870658427445131
- Model 2: Logistic Regression and TfidfVectorize
  - Train score 0.9159031181191846
  - Test score 0.8741877064024715
  - Best score: 0.8576603451949713
- Model 3: MultinomialNB and CountVectorizer (2nd Best Model)
  - Train score 0.9965551530648483
  - Test score 0.8732289336316182
  - Best score: 0.8592939839477236
- Model 4: MultinomialNB and TfidfVectorizer
  - Train score 0.907202216066482
  - Test score 0.8547991903696601
  - Best score: 0.850948220754315

Model 1 **(Logistic Regression and CountVectorize)** has the highest test score as well as Best_score followed by Model 3 **(MultinomialNB and CountVectorizer)**. So we would choose model 1 and would do futher analysis and prediction with model 1.

NYU

Model Evaluation

Workflow :

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

Model 1 (Best Model):

Accuracy: 88.35%
Precision: 84.99 %
Recall: 96.36 %
Specificity: 77.98 %
Misclassification Rate:14.63 %

VS

Base Accuracy: (Percent of class in the data)

Real News 0.56%
Fake News 0.43%

NYU

Modeling & Prediction

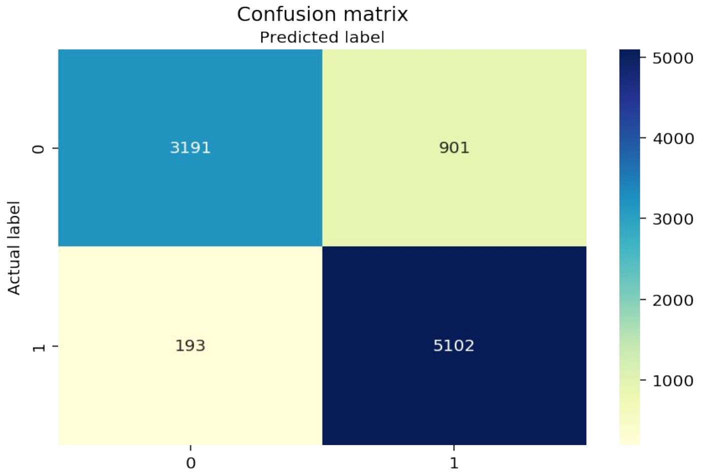
Workflow :

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

Model 1 Evaluation:

Accuracy: 88.35%
Precision: 84.99 %
Recall: 96.36 %
Specificity: 77.98 %
Misclassification Rate: 14.63 %

Confusion Matrix:



Real News : 0, Fake News : 1

Intuition:

We can see that Recall 96.36% is high, so our model is catching a lot of fake news, but at the same time Specificity is 77.98%, which is low and means our model is categorizing even real news as counterfeit at times. To deal with this, we would propose to add text sentiment as a feature and utilize the whole article. One should also carefully formulate the dataset such that it contains all types of news, and one or two entity doesn't take the entire scope of the dataset. For example, most of the news revolves around us election, president Trump, and Clinton, and that can affect how well the model would perform on other news.

Limitations:

1. The model utilizes only the Title of the article, not the whole article.
2. The model doesn’t take into consideration the sentiment of the text. i.e., if the text is representing hate or negativity etc.

NYU

Summary

Title: Fake News Detector

1) Problem Statement:

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?

2) Predictor and Target Variable(s):

- Predictor variable: Title
- Target Variable: Fake (1 for Fake or 0 for real.

3) Data source:

·

1. Reddit
2. The Guardian
3. Daily Mail
4. Gossipcop
5. Politifact
6. Washington University Fake News dataset

4) Data Preparation and EDA:

- Since I am working with the text, I need to use NLP methodology to scrutinize the text. Two of them would be.
  - Count Vectorizer
  - Term Frequency- Inverse Document Frequency

5) Modeling:

This project Falls under the supervised learning Classification problem, as the dataset contains the target labels and labels are discrete 0 or 1.

To name some:

- Logistic regression
- Multinomial Naive Bayes

6) Model Evaluation:

Our current goal is to increase the accuracy of the model as we want to focus on both telling the real news as real and fake as fake. So we are emphasizing on the Accuracy of the Model.

- Accuracy

Future Scope

Future Scope:

1. The first step, To gather more Fake news data. It is tough to get hands-on data that are fake as it is not readily available, and one cannot just assume that a given bad site would have all fake articles.
2. The second step would be to include the whole article and not just restrict ourselves with the article text.
3. The third step, Creating a new feature that has the sentiment of the article with each article. This feature may end up being very useful while classifying News as fake or Real.

Team Evaluation Report

Netid	Score
ksw352	4
asn419	4