Project 3

Classifying Real and Fake News

Content

- Intro / Background / Problem Statement
- Data Collection
- Data Cleaning and EDA
- Preprocessing and Modelling
  - LOGISTIC REGRESSION
  - NAIVE BAYES
- Evaluation + Insights
- Conclusion, Limitations and Future Outlook

Real or Fake? Can you tell?

9,000 NYC workers on unpaid leave for not complying with vaccine requirement. 91% did get at least one dose

cnn.com/2021/1...

582 Comments Share Save Hide Report 94% Upvoted

Trump Worried Biden Will Take Credit For 500,000 Covid Deaths He Made Possible

politics.theonion.com/trump-...

7 Comments Share Save Hide Report 97% Upvoted

Real vs Fake? How can we tell the difference?

r/news Posted by u/kiddenz 2 days ago

9,000 NYC workers on unpaid leave for not complying with vaccine requirement. 91% did get at least one dose

cnn.com/2021/1...

582 Comments Share Save Hide Report

94% Upvoted

r/TheOnion Posted by u/dwaxe 8 months ago

Trump Worried Biden Will Take Credit For 500,000 Covid Deaths He Made Possible

politics.theonion.com/trump-...

FAKE NEWS

7 Comments Share Save

97% Upvoted

Problem Statement

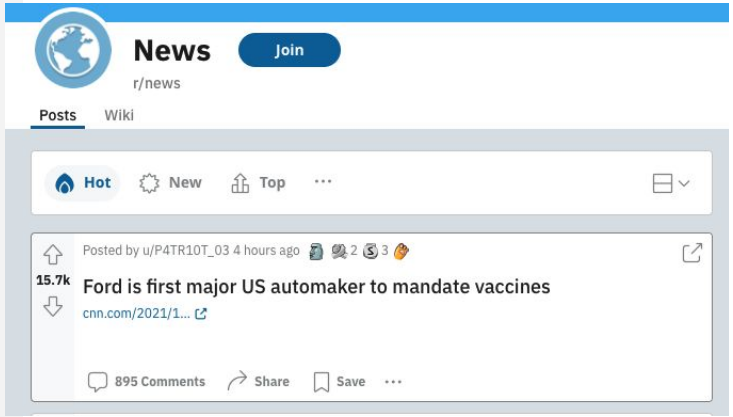
The increasing popularity and reliance on social media as a source of information worsens the problem of **misinformation** (**fake news**) today.

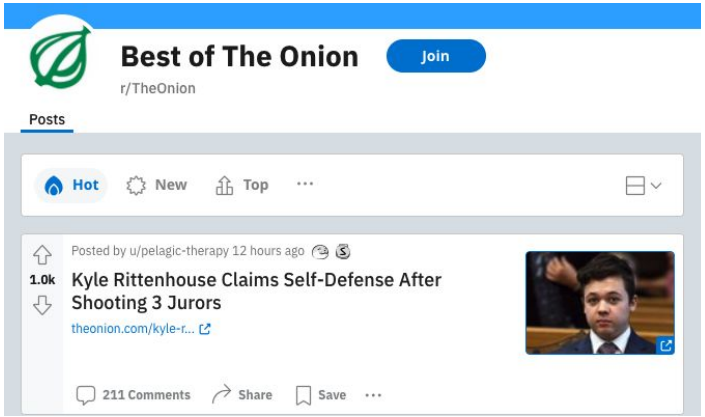
As a team of data scientist hired by a US government agency, we aim to analyse and **detect fake news** using **Natural Language Processing (NLP)** using data from the News and The Onion subreddit threads.

Why is fake news dangerous?

- Many people cannot tell the difference between real and fake news
- Creates confusion and misunderstanding about important socio-political issues
  - Cascades down to cause more dangerous widespread effects
    - Trump: used social media to spread misinformation and divide the people
    - Covid-19: Ivermectin as a cure

Background





- News articles current affairs in the US and the world
- Satirical, fake news

Data Collection I

Using **PushShift API**

- PushShift allows for **only 100** posts each retrieval
  - Created a **loop** to retrieve posts from the subreddits since
  - Set a **sleep** timer between loops so we wouldn't get blocked
  - Retrieved **10,000** posts from **each subreddit**

- Filter through to obtain posts **not removed** by moderators
  - 'removed_by_category': 'Nan'

- Posts were retrieved backwards from **25th Oct 2021**
  - UTC: 1635120000

Data Cleaning

دانلود آهنگ کردی گیس چنریا از ساسان ملکی

Volcán Monte Aso entra en erupción en Japón

↓ **Remove rows that aren't in English** > 2 non-alphanumeric characters

Data Cleaning

دانلود آهنگ کردی گیس چنریا از ساسان ملکی

Volcán Monte Aso entra en erupción en Japón

Remove rows that aren't in English

> 2 non-alphanumeric characters

https://edition.cnn.com/2021/09/16/us/florida....

Remove rows with links

anything containing "http"

Data Cleaning

دانلود آهنگ کردی گیس چنریا از ساسان ملکی

Volcán Monte Aso entra en erupción en Japón

Remove rows that aren't in English

> 2 non-alphanumeric characters

https://edition.cnn.com/2021/09/16/us/florida....

Remove rows with links

anything containing "http"

test

Ouch.

onion ice

Remove outliers in terms of length

< 5 words in length

Data Cleaning

دانلود آهنگ کردی گیس چنریا از ساسان ملکی

Volcán Monte Aso entra en erupción en Japón

Remove rows that aren't in English

> 2 non-alphanumeric characters

https://edition.cnn.com/2021/09/16/us/florida....

Remove rows with links

anything containing "http"

test

Ouch.

onion ice

Remove outliers in terms of length

< 5 words in length

Remove duplicate headlines

r/news **9336**

r/TheOnion **8872**

Data Cleaning

I am a happy person. I am filled with happiness. 我很快乐!

↓ **Remove punctuation and non-alphanumeric characters**

I am a happy person I am filled with happiness

↓ **Tokenize**

I, am, a, happy, person, I, am, filled, with, happiness.

↓ **Remove stop words**

happy, person, filled, happiness

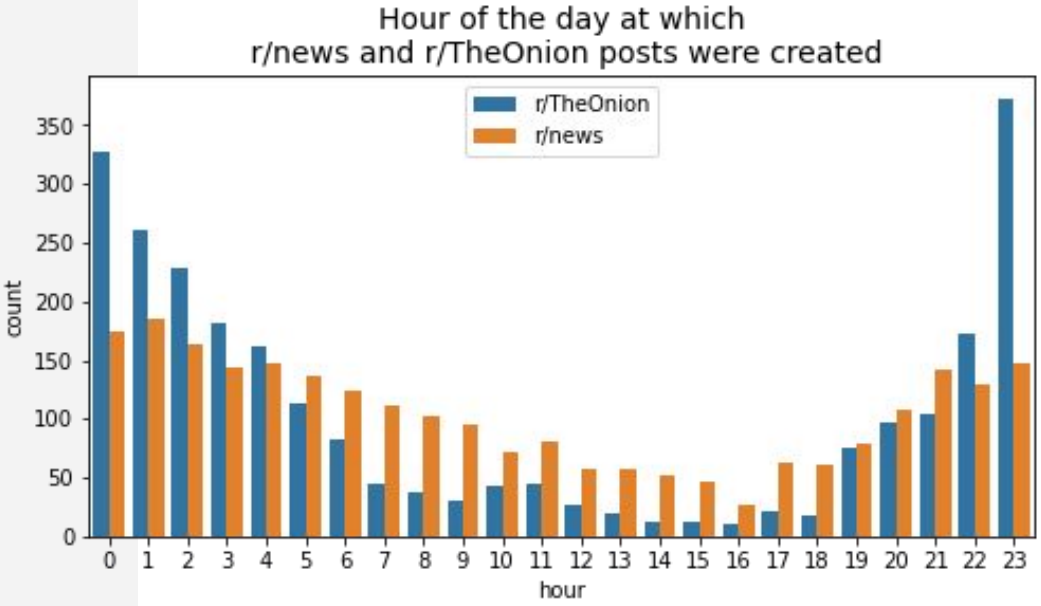
↓ **Lemmatizing**

happy, person, fill

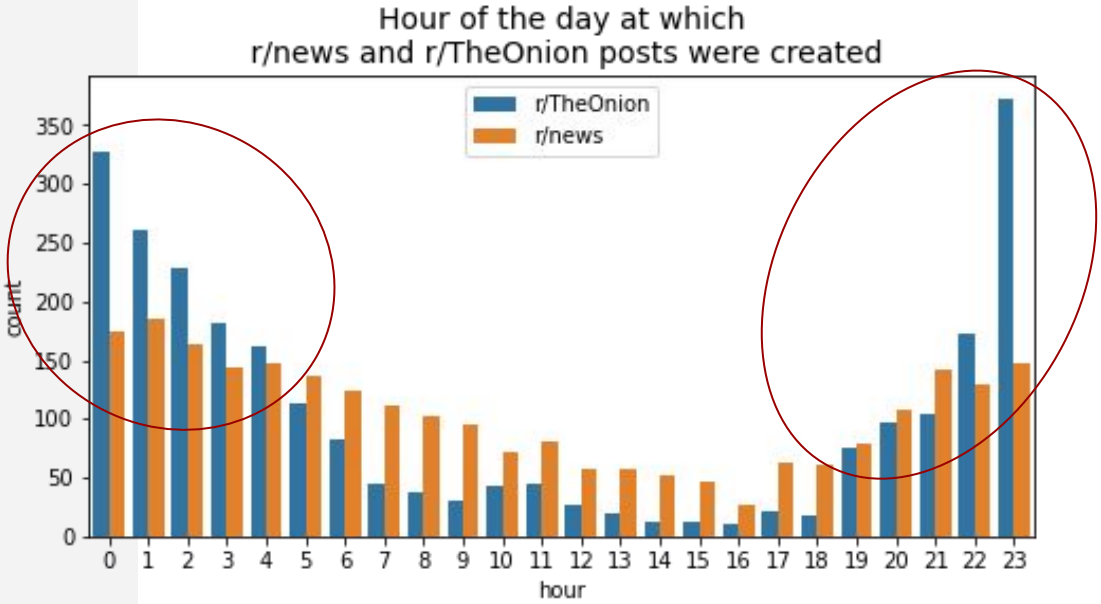
Exploratory Data Analysis

1. Hour of Posting
2. Length of Headline
3. Number of Comments
4. Month and Year of Posting
5. Headline

Time of posts

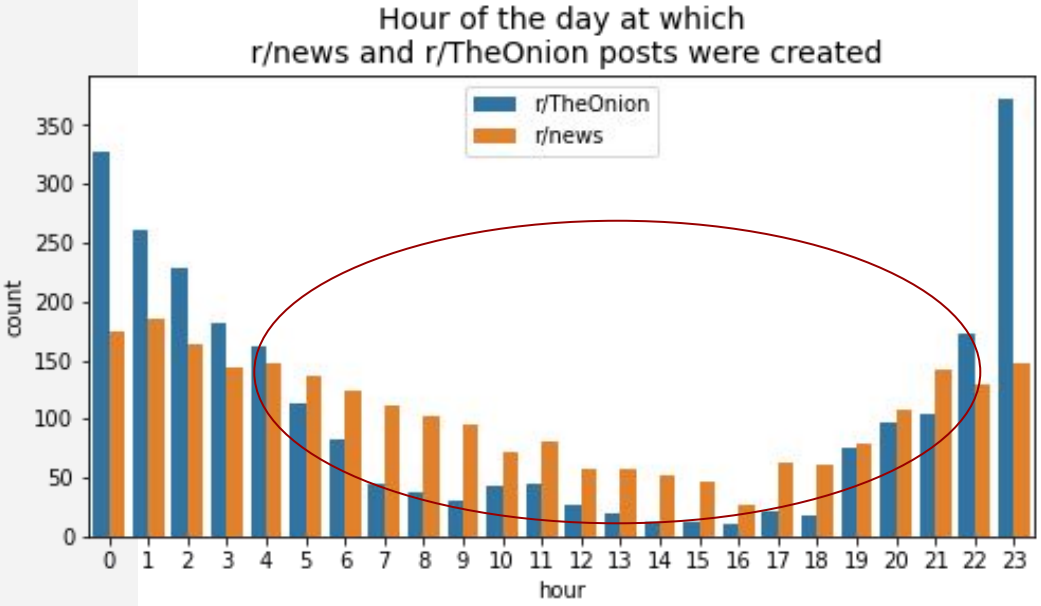


Time of posts



Between 10PM - 4AM is when the bulk of r/TheOnion posts are made.

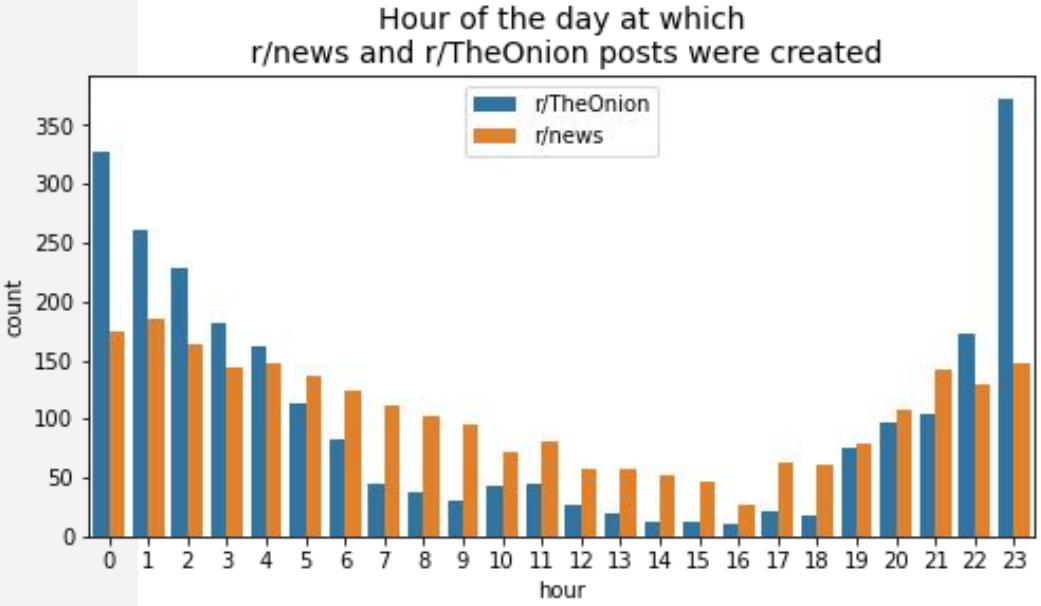
Time of posts



Between 10PM - 4AM is when the bulk of r/TheOnion posts are made.

Between 5AM - 9PM, the volume of r/news posts overtake r/TheOnion

Time of posts

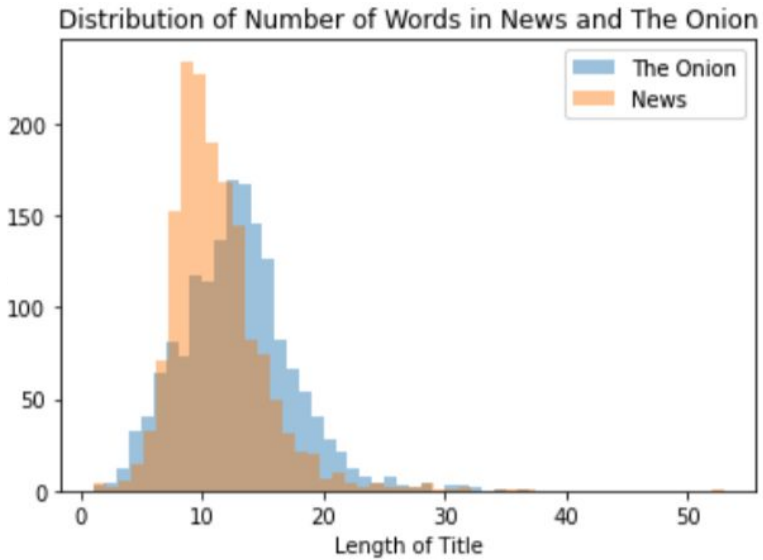


Between 10PM - 4AM is when the bulk of r/TheOnion posts are made.

Between 5AM - 9PM, the volume of r/news posts overtake r/TheOnion

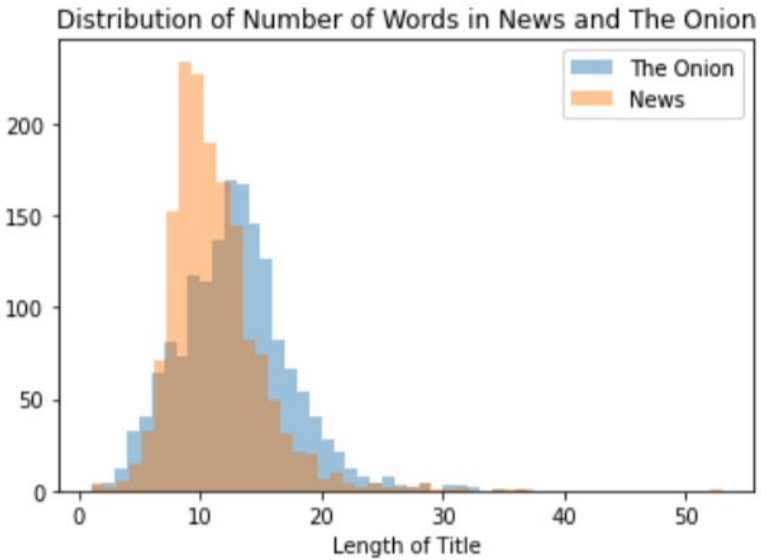
Posts on r/news were created more consistently through the day

Title length and Number of comments

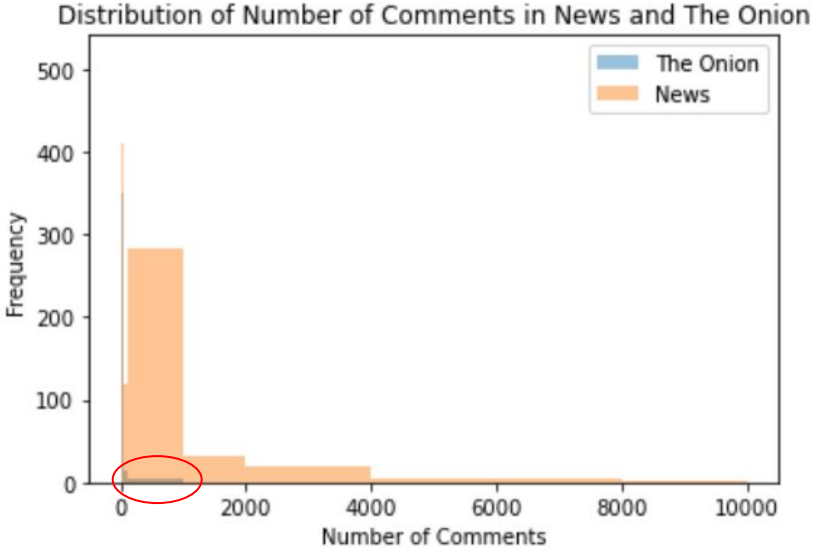


Distribution has a right skew, r/news titles tend to have less words than r/TheOnion

Title length and Number of comments

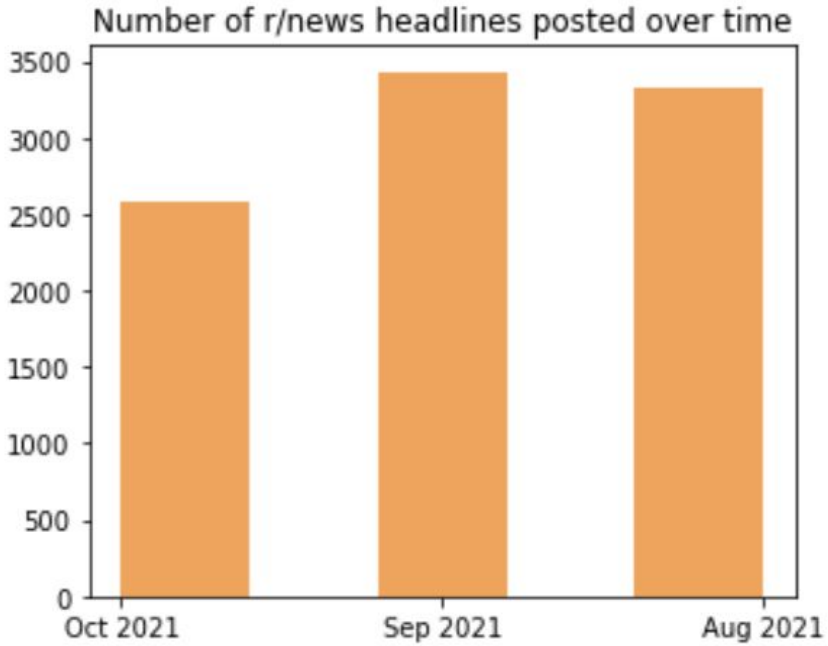


Distribution has a right skew, r/news titles tend to have less words than r/TheOnion



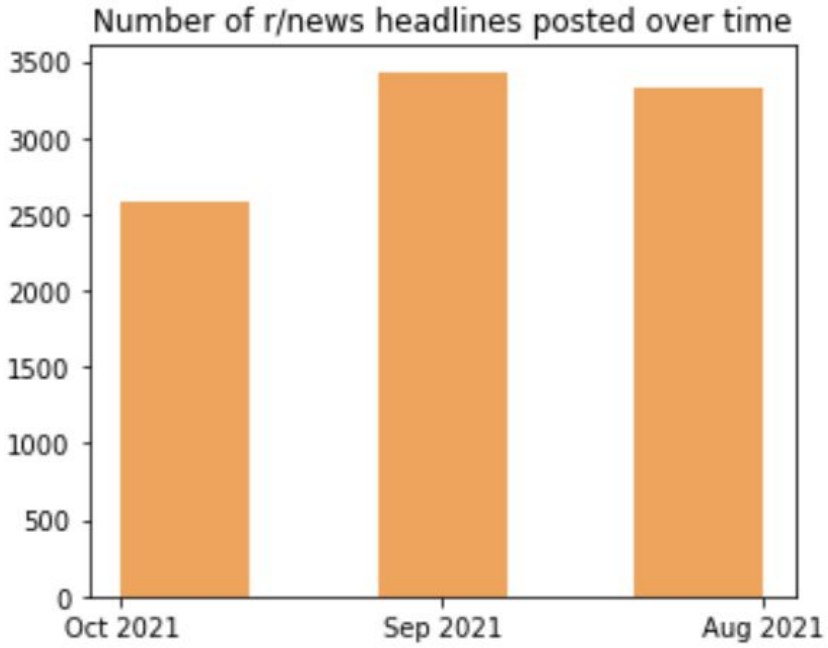
Posts from r/news generally receive more comments than posts from r/TheOnion

Month and year of posts



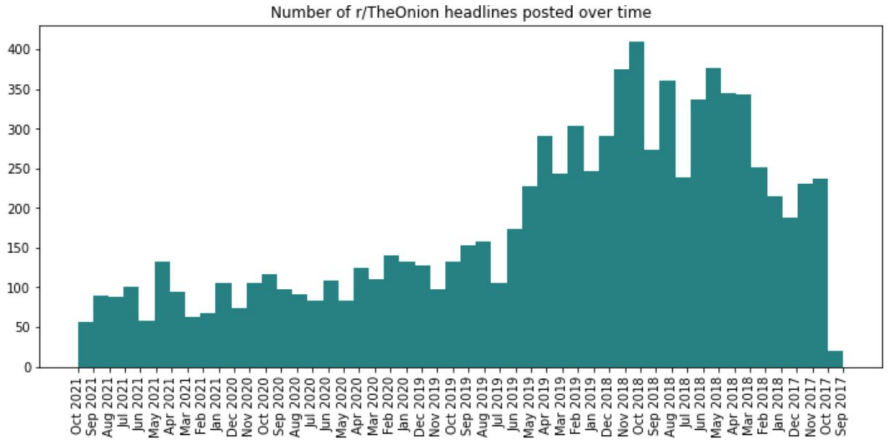
The r/news headlines were all created in the last three months from August to October

Month and year of posts



The r/news headlines were all created in the last three months from August to October

r/TheOnion headlines spanned back to 2017



Top words in r/news

Top words in r/news

Pandemic

Top words in r/news

Pandemic

Violent crimes

Top words in r/news

Pandemic

Violent crimes

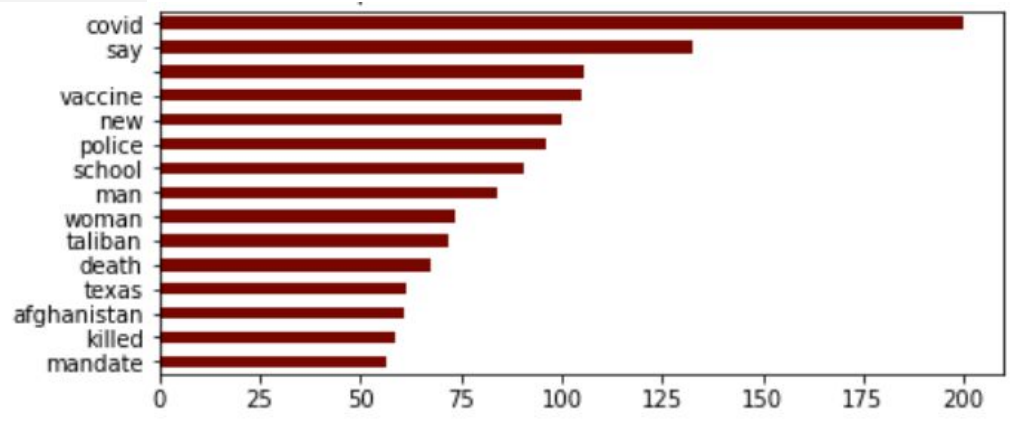
Recent politics news

Top words in r/news

Pandemic

Violent crimes

Humanitarian crises



Top words in r/TheOnion

Use of more general words

Top words in r/TheOnion

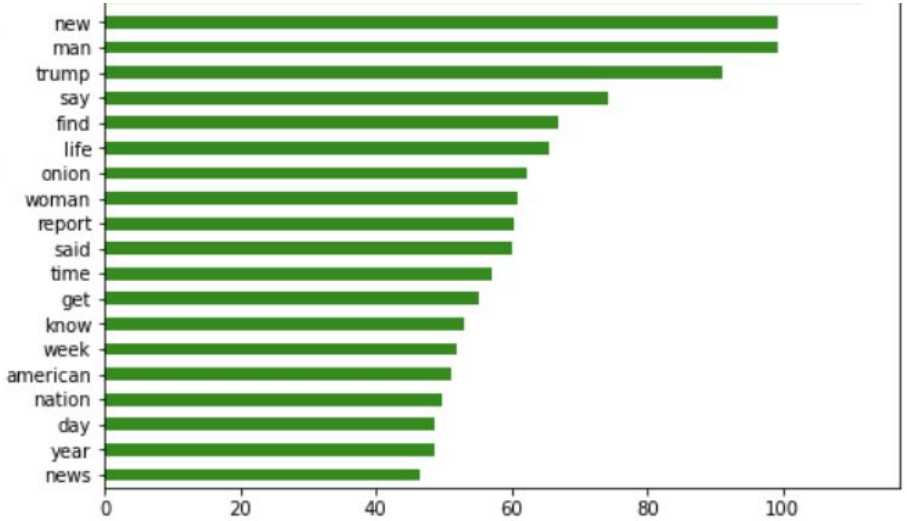
Use of more general words

(Except for Trump)

Top words in r/TheOnion

Use of more general words

(Except for Trump)



Preprocessing and Modeling – **Overview**

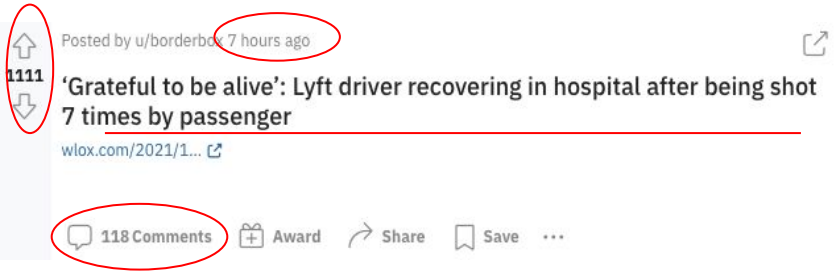
Detecting fake news (2 approaches)

Non-text features

- Score
- Comments
- Length of post
- Time

Text features

- Headline text



Preprocessing and Modeling - **Non-text features**

- **Preprocessing:**
  - One-hot encoded `hours` from time (epoch)
  - Scaling of data

num_comments	score	num_char	title	created_utc
9093	67198	49	FBI raids New York City po..	1633449271
17764	54117	74	Capitol Police officer who..	1629487987
4675	51147	63	Third Sandy Hook parent wi...	1633475492

num_comments	score	num_char	title	hour_0	hour_1	hour_2	hour_3	hour_4	hour_5	hour_6
0.511878	1.000000	0.163823	FBI raids New Yo...	0	0	0	0	0	0	0
1.000000	0.805334	0.249147	Capitol Police o...	0	0	0	1	0	0	0
0.263173	0.761135	0.211604	Third Sandy Hook...	0	0	0	0	0	0	0

Preprocessing and Modeling - **Non-text features**

- **Modeling:**
  - Logistic Regression
  - K-Nearest Neighbors
  - Naive Bayes

- Hyperparameter tuning thru GridSearchCV

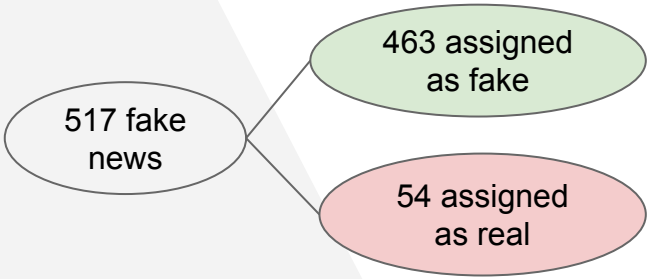
Best model: Logistic Regression

	Train Accuracy	Test Accuracy
Before hyperparameter tuning	0.8445	0.852
After hyperparameter tuning	**0.906**	**0.904**

features	coef
num_comments	77.436814
score	20.938244
num_char	0.489549
hour_23	0.148621
hour_21	0.140427
hour_1	0.129548
hour_17	0.121468

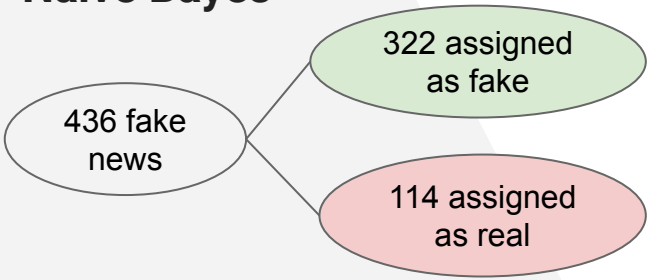
Comparison (Non-text features)

- **Logistic Regression**

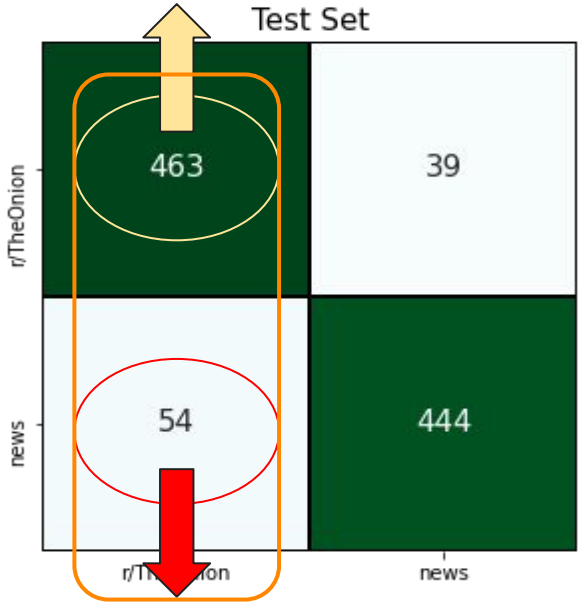


10%

- **Naive Bayes**



26% !



Preprocessing and Modeling - **Text features (NLP)**

- **Preprocessing:**
  - Conversion of textual data into vectorized forms:
    - Count Vectorizer
    - TF-IDF Vectorizer

Illegal Immigrants Returning To Mexico For American Jobs
News: Box Office Failure: The $450 Million 'The Last Jedi' Made On Its Opening Weekend Is Less Than A Third Of What The Cast And Crew Spent On Paper Towels
News: Accessibility FTW! Tic Tac Is Making Its Breath Mints 500 Times Larger For The Visually Impaired
Middle-Aged Woman Angrily Demanding Price Check Was Once Carefree Youth, Onlookers Speculate

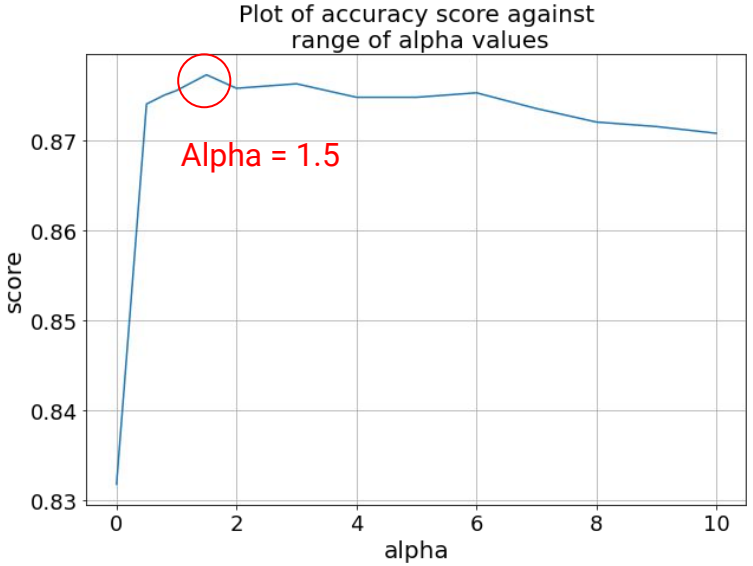
columbus
loan
whenever
senseless
halloween
gunned
employee
settling
play
antigovernment

activist	afghan	afterlife	allegedly	allows	angeles	animatronic
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.377964
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.447214	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.383687	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.395751	0.000000

Preprocessing and Modeling - **Text features (NLP)**

- **Modeling:**
  - Logistic Regression
  - K-Nearest Neighbors
  - Naive Bayes

- Hyperparameter tuning thru GridSearchCV
  - Optimized alpha value of 1.5

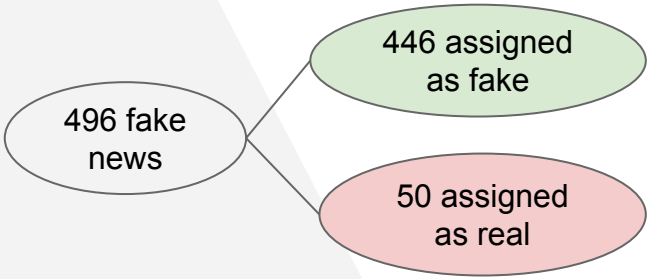


Best model: Multinomial Naive Bayes

	Train Accuracy	Test Accuracy
Before hyperparameter tuning	0.875	0.829
After hyperparameter tuning	**0.877**	**0.904**

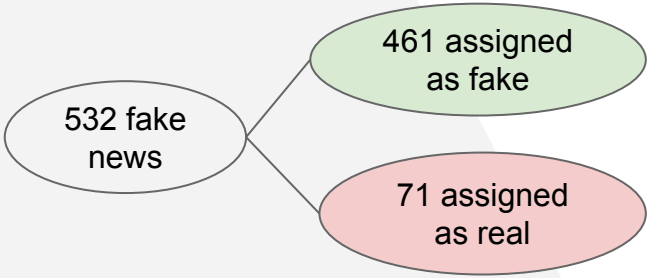
Comparison (Text features)

- **Naive Bayes**

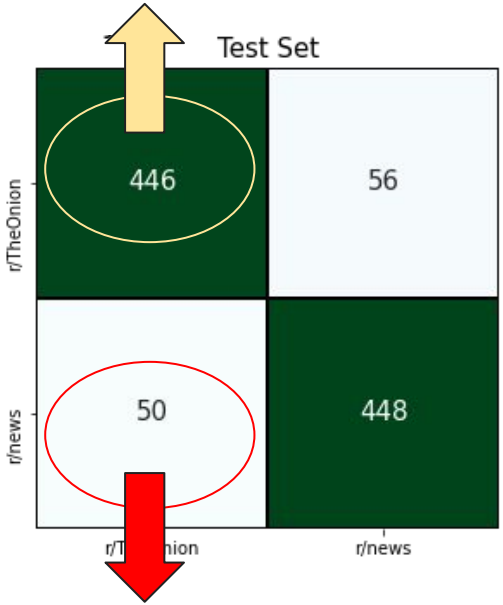


10%

- **Logistic Regression**



13%



Evaluation: Modeling for Non-Text Features

	Logistic Regression	Naive Bayes Multinomial	K-Nearest Neighbour
Train Accuracy	0.930	0.704	0.857
Mean CV Accuracy	0.918	0.693	0.725
Test Accuracy	**0.907**	**0.706**	**0.743**
Precision	**0.919**	**0.681**	**0.722**
Recall	0.892	0.771	0.787
F1-Score	0.905	0.723	0.753

Precision = True Positive / (True Positive + False Positive)

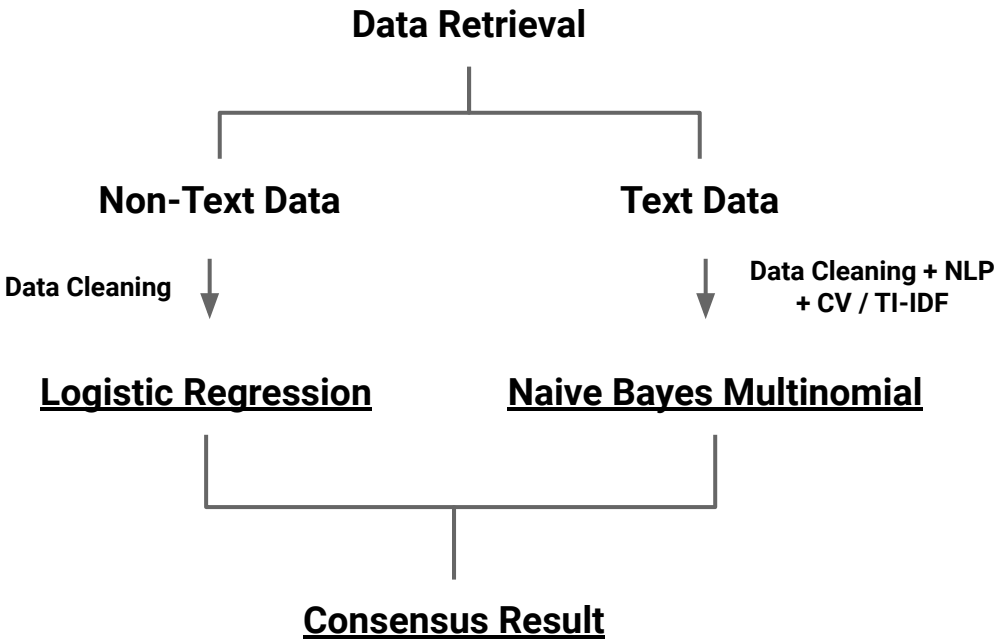
Evaluation: Modeling for Text Features

	Logistic Regression	Naive Bayes Multinomial	K-nearest neighbour
Train Accuracy	0.995	0.967	0.865
Mean CV Accuracy	0.876	0.877	0.774
Test Accuracy	**0.888**	**0.894**	**0.78**
Precision	**0.905**	**0.889**	**0.713**
Recall	0.865	0.90	0.933
F1-Score	0.885	0.894	0.809

Evaluation: Combining both models

What have we done so far and why?

Correct prediction of real news by Log Reg (Non-text data):	91.9%
Correct prediction of real news by Naive Bayes (Text data):	89%
Correct prediction by consensus result :	97.8%



Another perspective from bayesian stats: The probability that a real news is real GIVEN that both models predicted that the post is real news

Extra explanation of consensus result

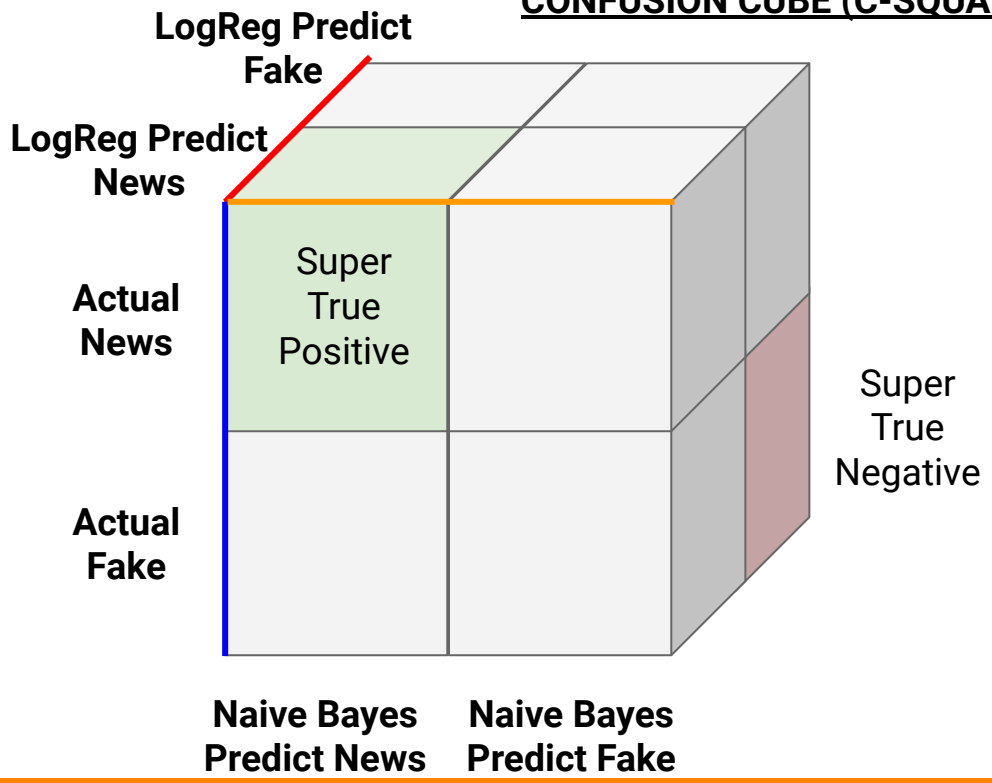
Model Results Confusion Matrix

	Predicted News	Predicted Fake
Actual News	TP	FN
Actual Fake	FP	TN

Positive Predictive Value = 0.978

Negative Predictive Value = 0.988

CONFUSION CUBE (C-SQUARED)



Evaluation: Illustration of both models in practice

An illustration

Evaluation: Illustration of both models in practice

An illustration

Logistic Regression using non-text features: **REAL NEWS**

Naive Bayes using Subreddit title (text) features: **REAL NEWS**

Evaluation: Illustration of both models in practice

An illustration

Logistic Regression using non-text features: **REAL NEWS**

Naive Bayes using Subreddit title (text) features: **REAL NEWS**

Conclusion

When non-text data is used, such as time the post was created, post score and number of comments, Logistic Regression far outperforms the rest of the models.

	Logistic Regression	Naive Bayes Multinomial	K-Nearest Neighbour
Train Accuracy	0.930	0.704	0.857
Mean CV Accuracy	0.918	0.693	0.725
Test Accuracy	**0.907**	**0.706**	**0.743**
Precision	**0.919**	**0.681**	**0.722**
Recall	0.892	0.771	0.787
F1-Score	0.905	0.723	0.753

Conclusion

Both Logistic Regression and Naive Bayes perform quite similarly when classifying text.

	Logistic Regression	Naive Bayes Multinomial
Train Accuracy	0.995	0.967
Mean CV Accuracy	0.876	0.877
Test Accuracy	**0.888**	**0.894**
Precision	**0.905**	**0.889**
Recall	0.865	0.90
F1-Score	0.885	0.894

Recommendations

It was interesting to note that Logistic Regression on non-text features had better scores overall when predicting whether a title was legit or not.

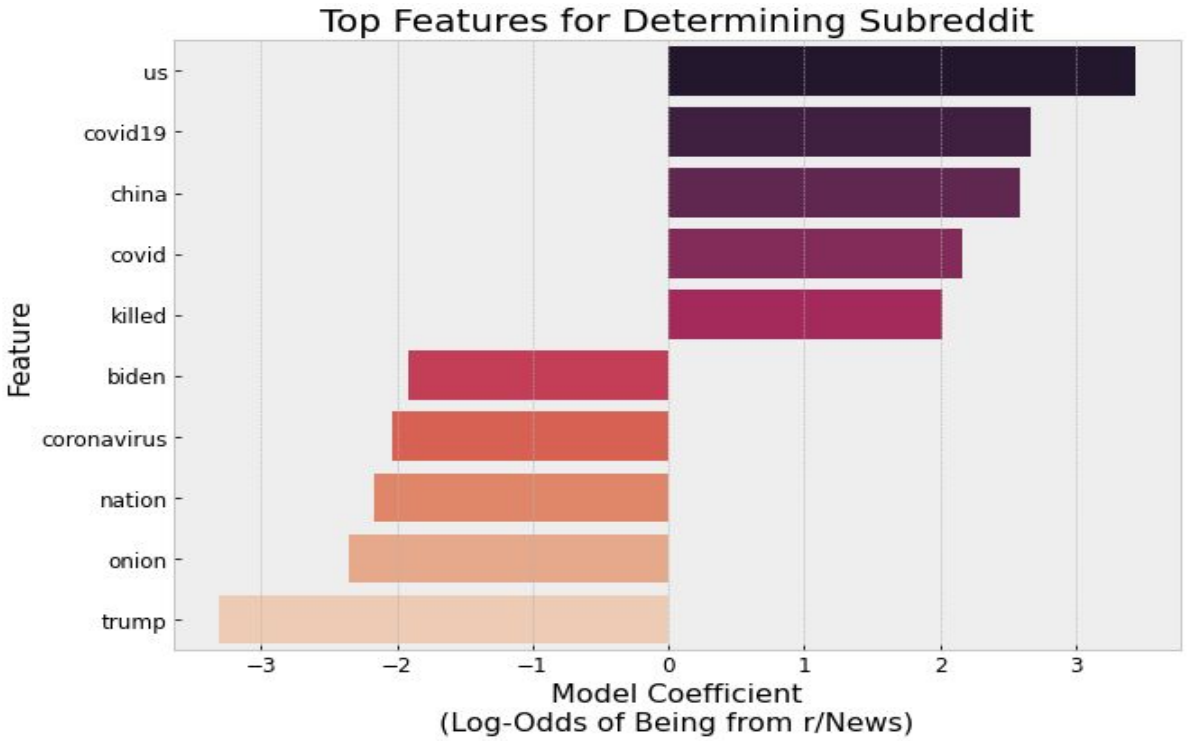
Recommendations

It was interesting to note that Logistic Regression on non-text features had better scores overall when predicting whether a title was legit or not.

However;

Our group recommends the textual count vectorized Naive Bayes model.

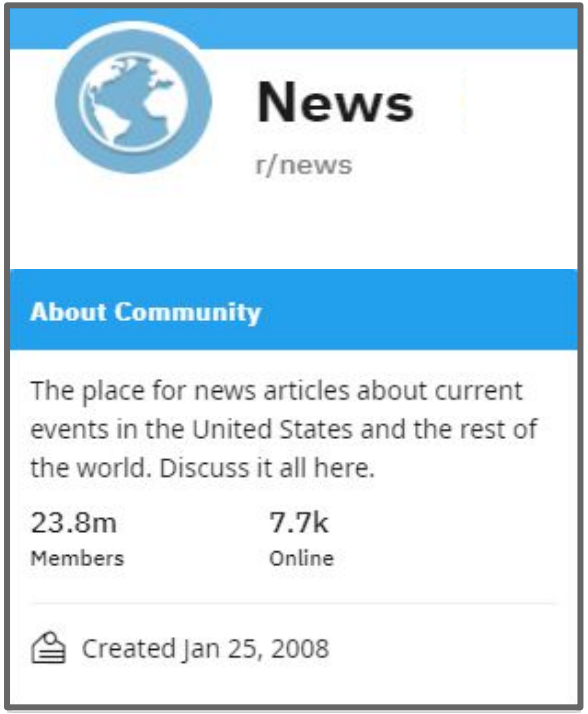
Limitations

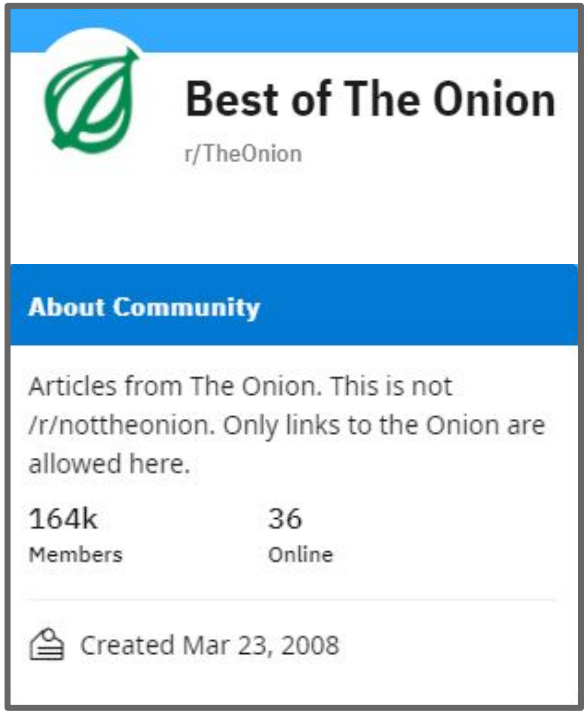


Misclassified r/TheOnion posts

D.C Police Preemptively Deploy 3 Officers For Inauguration Day
Covid Denier Struggling To Protest State’s Incoherent, Constantly Changing Coronavirus Policies
Unvaccinated Mom Wants To Know If You’re Coming Home For Covid This Year
U.N. Court Orders U.S. To Ease Sanctions Against Iran

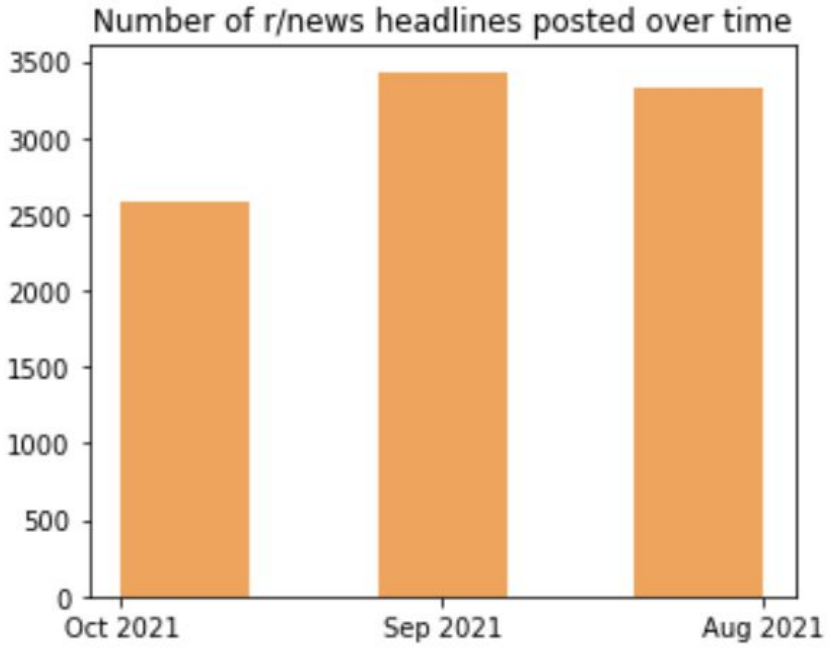
Subreddit Members





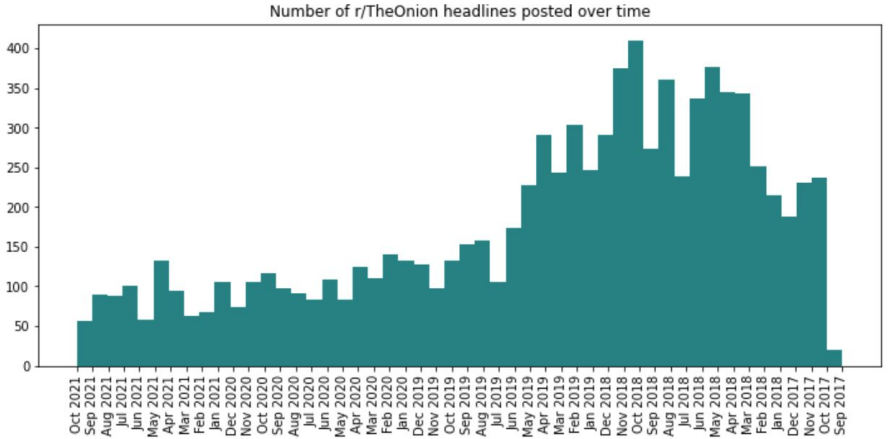
Limitations

Limitations



The r/news headlines were all created in the last three months from August to October

r/TheOnion headlines spanned back to 2017

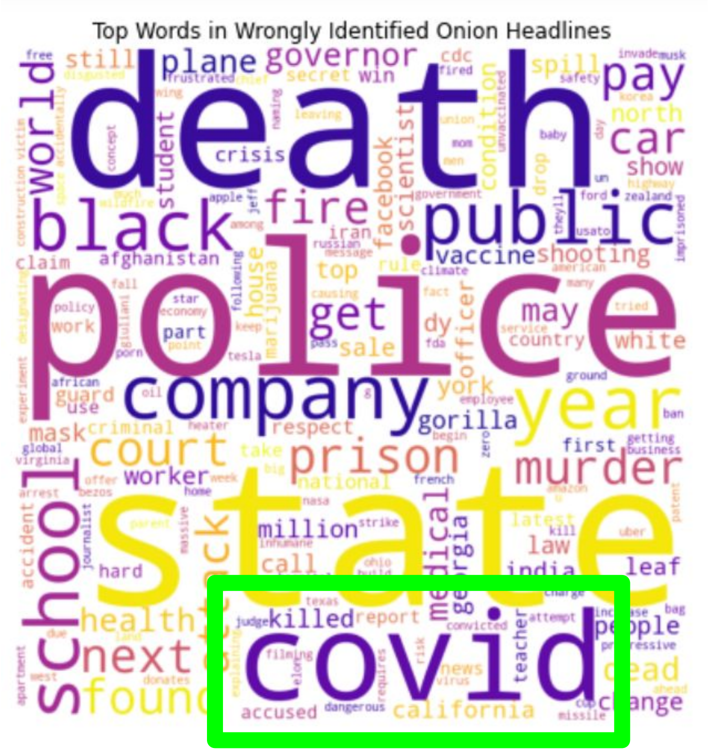



Top Words in Wrongly Identified Onion Headlines
death
police
state
covid
company
public
black
school
year
prison
court
murder
fire
world
plane
governor
pay
car
get
attack
next
found
health
medical
georgia
million
killed
dead
people
change
california
vaccine
shooting
gorilla
india
leaf
law
white
may
afghanistan
student
scientist
facebook
marijuana
officer


school
covid
police
vaccine
taliban
death
texas
mandate
florida
people
year
afghanistan
killed
china
child
state
case
california
officer
mask
shot
shooting
dy
dead
get
student
court
found
afghan
report
worker
charged
hospital
american
world
million
official
vaccination
county
former
attack


Top Words in Wrongly Identified Onion Headlines
death
police
state
covid
company
public
black
year
school
prison
murder
court
plane
governor
pay
car
world
get
may
health
next
found
attack
medical
georgia
india
million
killed
people
dead
change
california
vaccine
shooting
afghanistan
white
sale
officer
gorilla
worker
law
leaf
call
hard
accused
crisis
student
still
north
show


school
covid
police
vaccine
taliban
death
texas
mandate
state
case
officer
california
florida
people
year
afghanistan
killed
child
mask
shooting
china
dy
get
student
hospital
dead
court
found
american
world
million
official
vaccination
county
former
arrested
vaccinated
unvaccinated
afghan


Top Words in Wrongly Identified Onion Headlines
death
police
state
public
black
company
year
prison
court
murder
school
fire
world
get
may
car
pay
governor
plane
covid
health
next
found
killed
people
california
million


school
covid
police
vaccine
taliban
death
officer
state
california
case
texas
florida
mask
shot
afghanistan
killed
shooting
china
people
court
found
worker
american
world million
official
vaccination
county


death
police
state
covid
company
black
public
school
year
prison
court
murder
found
next
health
attack
medical
georgia
million
killed
people
world
plane
governor
still
student
crisis
car
north
show
vaccine
shooting
afghanistan
get
may
white
sale
top
house
marijuana
officer
york
gorilla
mask
criminal
worker
first
law
india
leaf
call
hard
report
dead
change
california
accused
national
respect
take
latest
claim
work
part
guard
use
condition
secret
win
spill
bay


school
covid
police
vaccine
taliban
death
texas
mandate
florida
people
year
child
state
california
case
officer
afghan
afghanistan
killed
china
dy
get
student
dead
shot
mask
shooting
report
court
found
worker
hospital
black
fire
record
judge
first
could
kabul
arrested
vaccinated
american
world
million
official
vaccination
county
former
attack
charged
yearold
day
video
facebook
murder
city
doctor
employee
ban
high
two
drug
face
india
time
find
family
call
federal


Top Words in Wrongly Identified Onion Headlines
death
police
state
covid
public
black
company
school
year
prison
court
murder
attack


school
covid
police
vaccine
taliban
death
afghanistan
killed
florida
people
year
texas
mask
shooting
state

Future Outlook

- Compare date and headline keywords.

Future Outlook

- Compare date and headline keywords.

- Look into other legit news sources to see if the model is able to pull its weight on different headlines from other places.

Future Outlook

- Compare date and headline keywords.
- Look into other legit news sources to see if the model is able to pull its weight on different headlines from other places.
- Apply models to r/NotTheOnion. (real news that sounds like fake news)

Future Outlook

- Compare date and headline keywords.
- Look into other legit news sources to see if the model is able to pull its weight on different headlines from other places.
- Apply models to r/NotTheOnion. (real news that sounds like fake news)
- Perform sentiment analysis.

Thank you!
Any Questions?

Requirements

- Gather and prepare your data using the `requests` library.
- **Create and compare two models**. One of these must be a Naive Bayes classifier, however the other can be a classifier of your choosing: logistic regression, KNN, SVM, etc.
- A Jupyter Notebook with your analysis for a peer audience of data scientists.
- An executive summary of your results.
- A short presentation outlining your process and findings for a semi-technical audience.

Pro Tip: You can find a good example executive summary here.

Presentation

- Is the problem statement clearly presented?
- Does a strong narrative run through the presentation building toward a final conclusion?
- Are the conclusions/recommendations clearly stated?
- Is the level of technicality appropriate for the intended audience?
- Is the student substantially over or under time?
- Does the student appropriately pace their presentation?
- Does the student deliver their message with clarity and volume?
- Are appropriate visualizations generated for the intended audience?
- Are visualizations necessary and useful for supporting conclusions/explaining findings?

For Project 3 the evaluation categories are as follows:

The Data Science Process

- Problem Statement
- Data Collection
- Data Cleaning & EDA
- Preprocessing & Modeling
- Evaluation and Conceptual Understanding
- Conclusion and Recommendations

Organization and Professionalism

- Organization
- Visualizations
- Python Syntax and Control Flow
- Presentation