

The background of the slide features a dark, blurred image of a laptop on a wooden desk, with a black pen resting on the laptop's lid. A large, solid orange triangle is positioned on the right side of the slide, pointing towards the center. The text is overlaid on the dark area of the laptop.

# 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/2021/10/26/health/nyc-workers-vaccine-requirement/index.html](https://www.cnn.com/2021/10/26/health/nyc-workers-vaccine-requirement/index.html)

 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-worried-biden-will-take-credit-for-500-000-covid-deaths-he-made-possible/](https://www.politics.theonion.com/trump-worried-biden-will-take-credit-for-500-000-covid-deaths-he-made-possible/)

 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...](https://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-...](https://politics.theonion.com/trump-...)

 7 Comments  Share  Save

**FAKE NEWS**



97% Upvoted

# Problem Statement

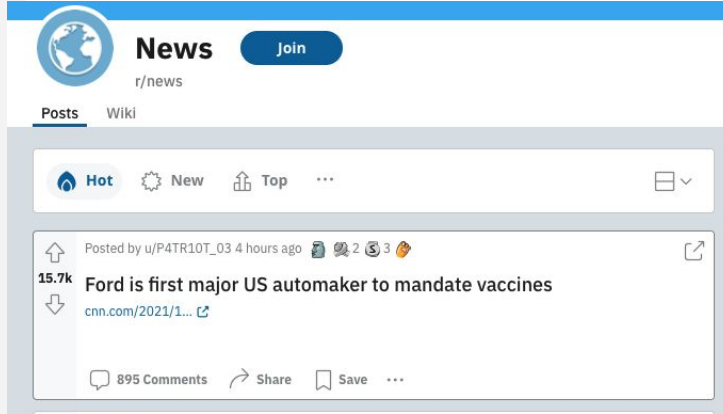
The increasing popularity and reliance on social media as a source of information worsens the problem of **misinformation (fake news)** today.

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.

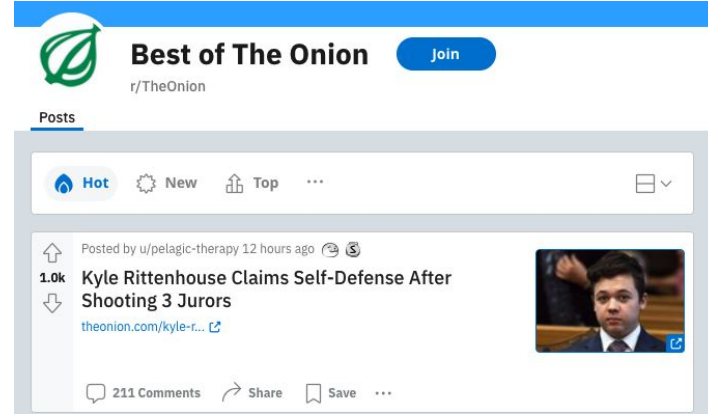
# Why is fake news dangerous?

- ▶ 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 headlines 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 headlines 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 headlines 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**

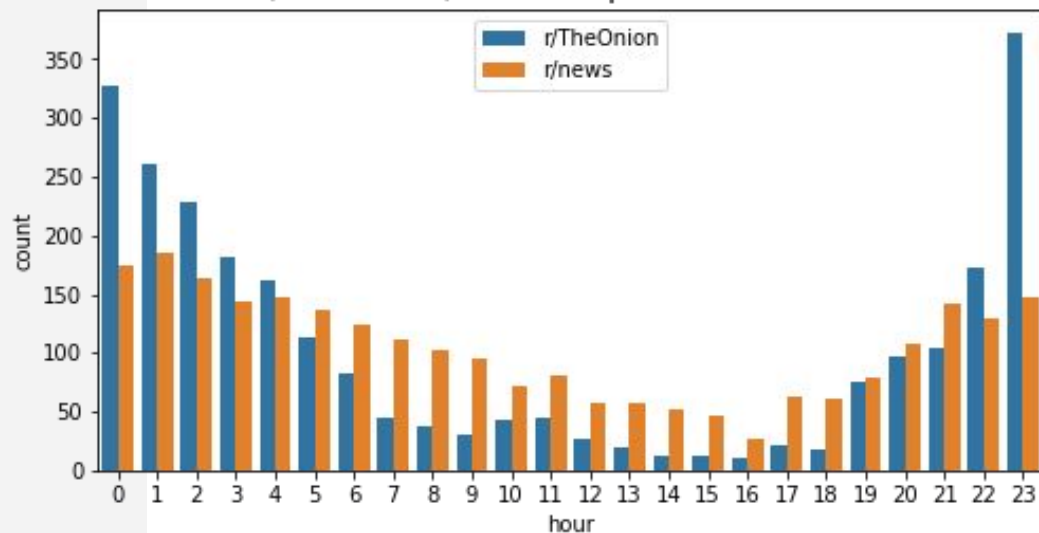
happy, person, fill

# Exploratory Data Analysis

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

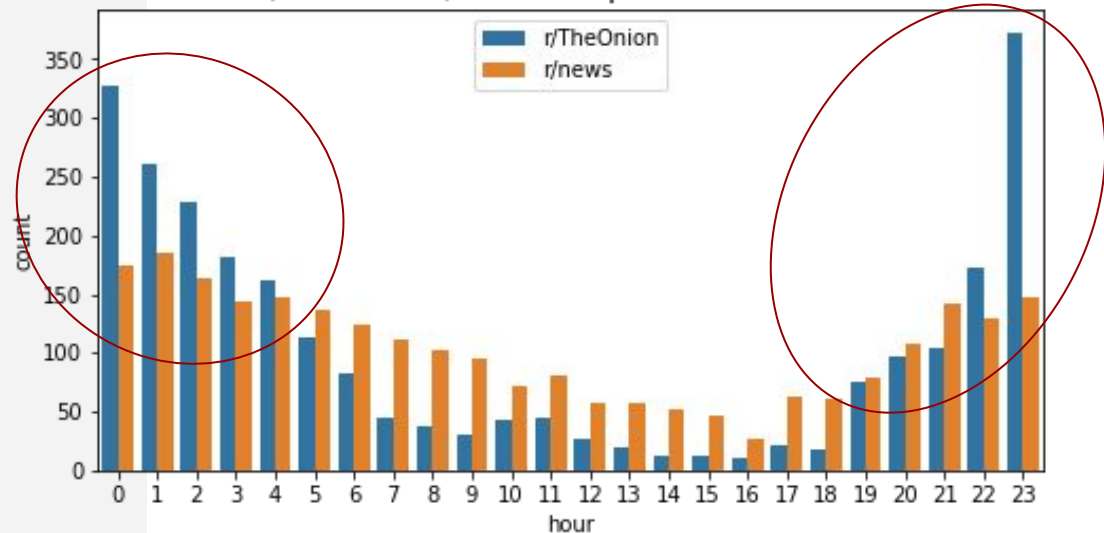
# Time of posts

Hour of the day at which  
r/news and r/TheOnion posts were created



# Time of posts

Hour of the day at which  
r/news and r/TheOnion posts were created

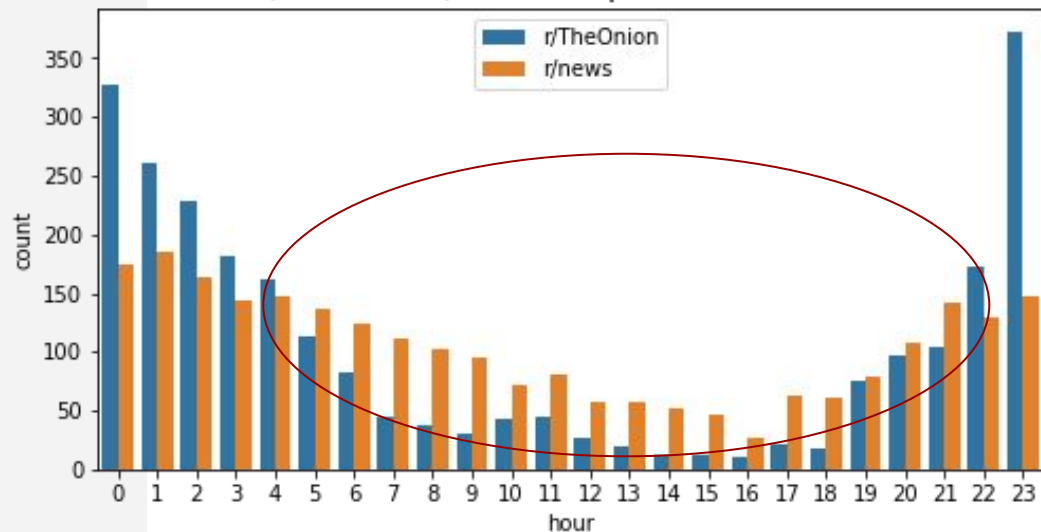


Between 10PM - 4AM is when the bulk of r/TheOnion posts are made.



# Time of posts

Hour of the day at which  
r/news and r/TheOnion posts were created

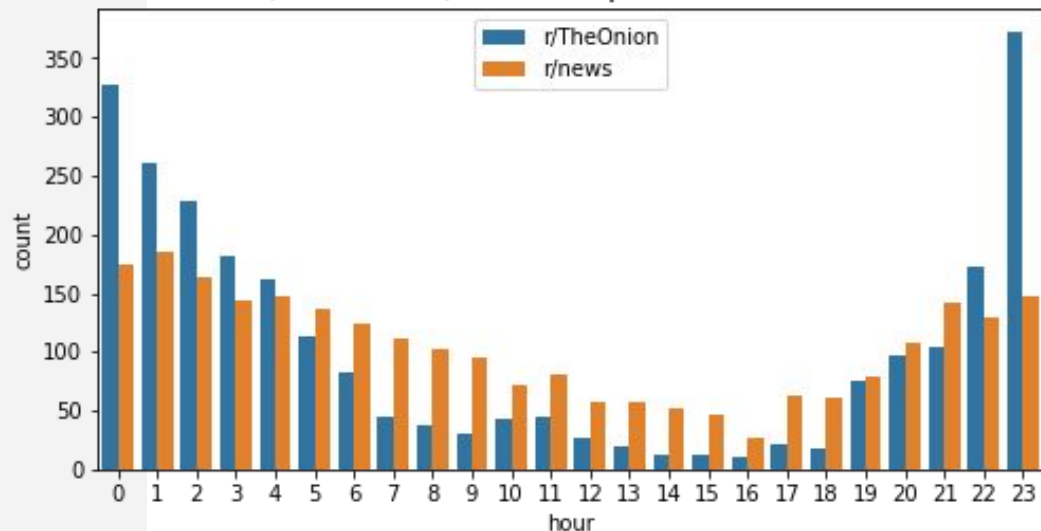


Between 10PM - 4AM is when the bulk of r/TheOnion posts are made.

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

# Time of posts

Hour of the day at which  
r/news and r/TheOnion posts were created

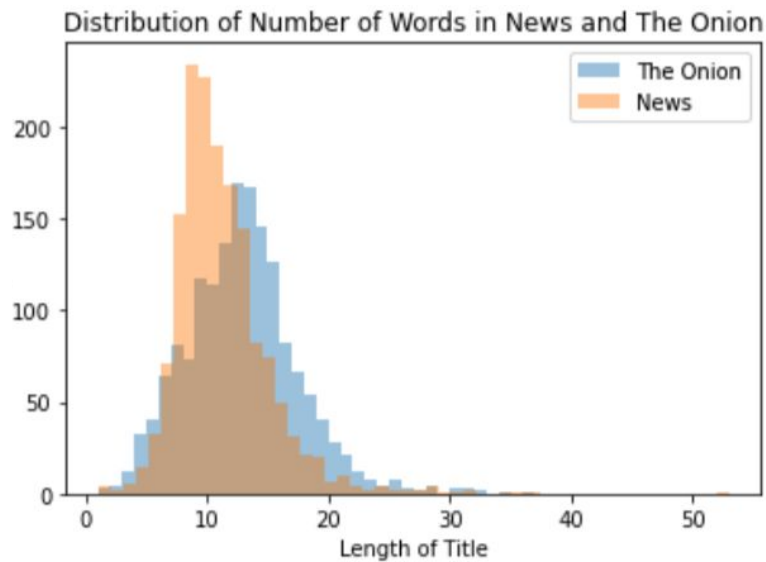


Between 10PM - 4AM is when the bulk of r/TheOnion posts are made.

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

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

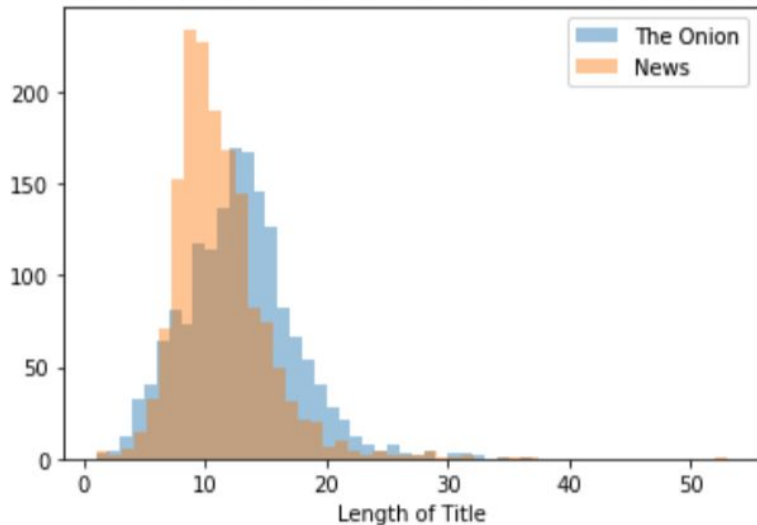
# Title length and Number of comments



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

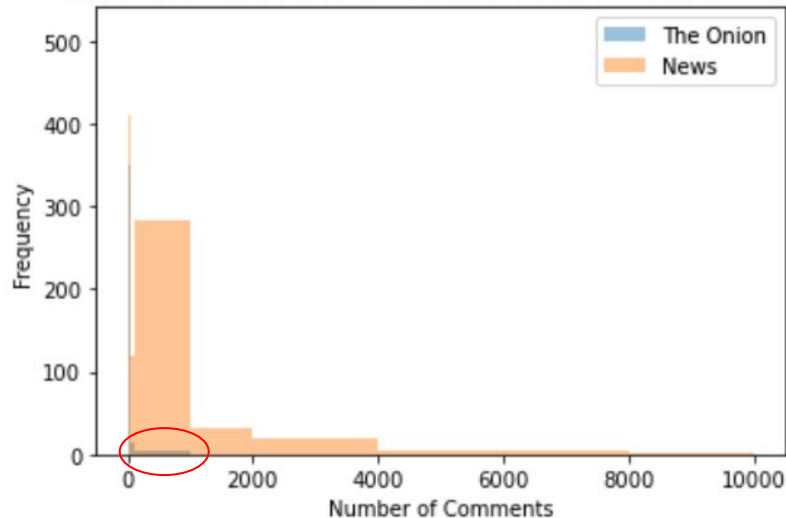
# Title length and Number of comments

Distribution of Number of Words in News and The Onion



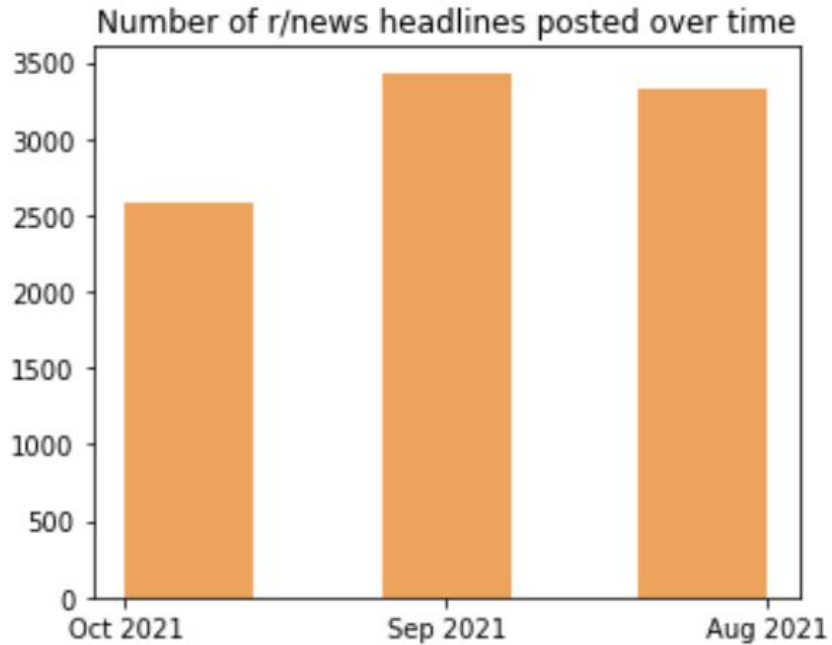
Distribution has a right skew, r/news titles tend to have less words than r/TheOnion

Distribution of Number of Comments in News and The Onion



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

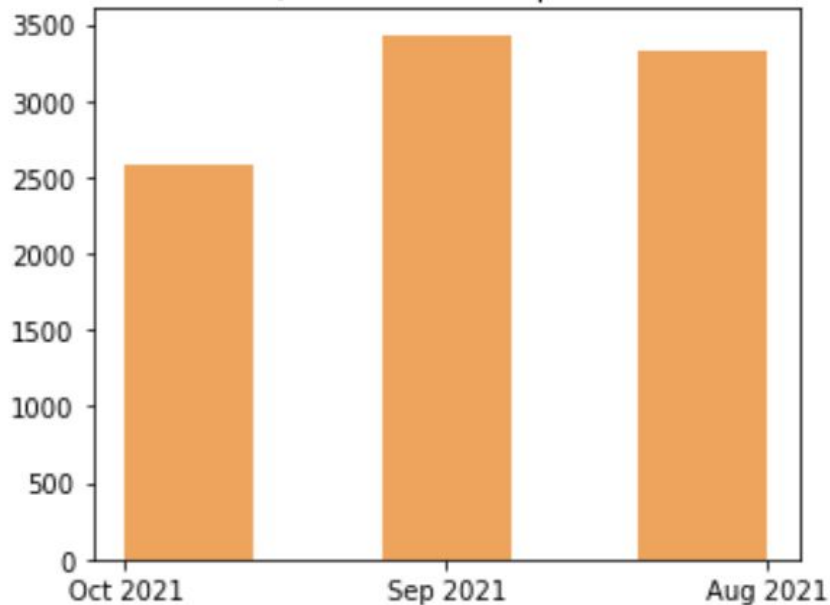
# Month and year of posts



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

# Month and year of posts

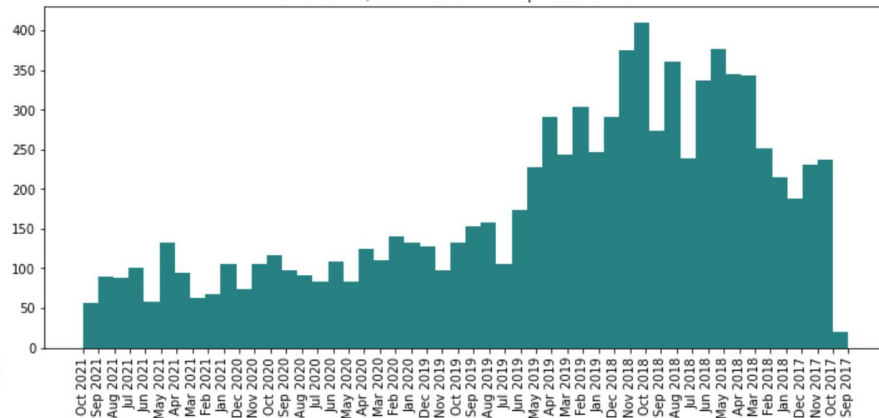
Number of r/news headlines posted over time



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

r/TheOnion headlines spanned back to 2017

Number of r/TheOnion headlines posted over time



# Top words in r/news



# Top words in r/news

# Pandemic





# Top words in r/news

Pandemic

Violent crimes



# Top words in r/news

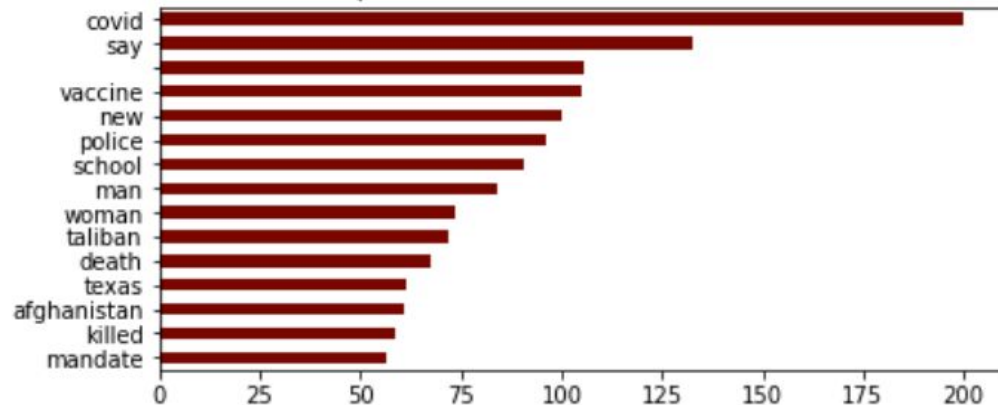
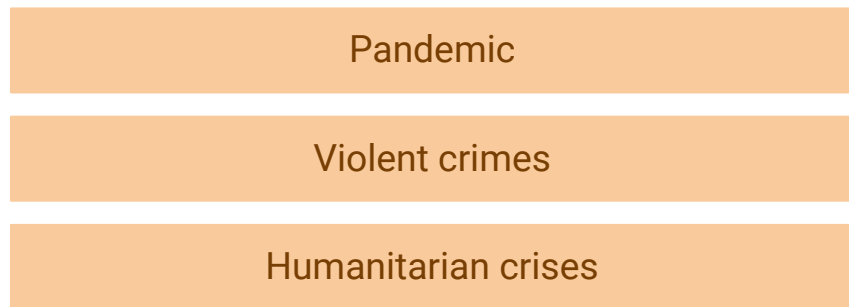
# Pandemic

## Violent crimes

## Recent politics news



# Top words in r/news



### Top Words in The Onion Headlines



## Use of more general words

### Top Words in The Onion Headlines



## Use of more general words

(Except for Trump)

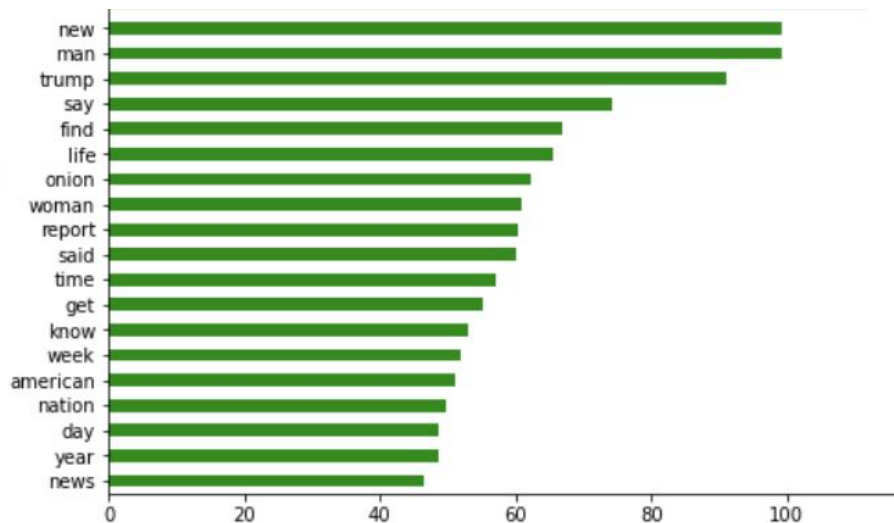


# Top words in r/The0nion



## Use of more general words

(Except for Trump)



# Preprocessing and Modeling – Overview

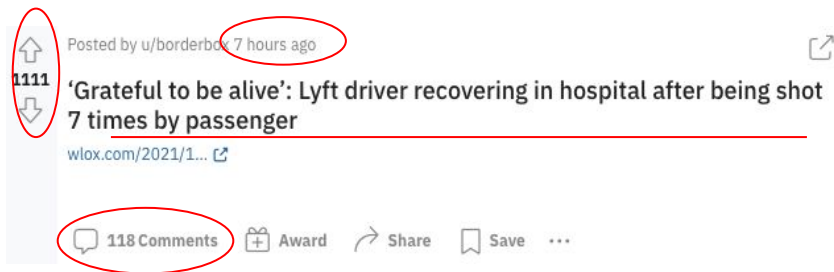
## Detecting fake news (2 approaches)

### 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

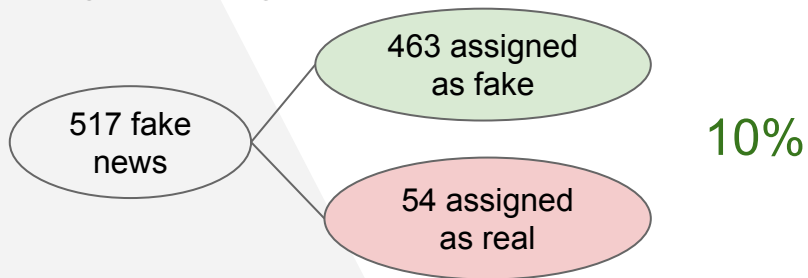
## **Best model: Logistic Regression**

	Train Accuracy	Test Accuracy
Before hyperparameter tuning	0.8445	0.852
After hyperparameter tuning	<b>0.906</b>	<b>0.904</b>

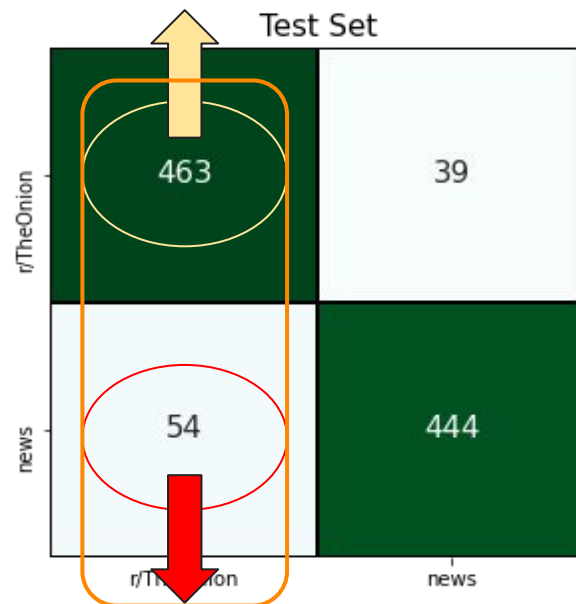
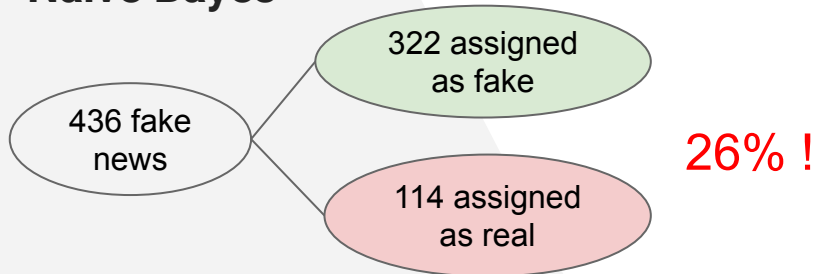
features	coef
num_comments	77.436814
score	20.938244
num_char	0.489549
hour_23	0.148621
hour_21	0.140427
hour_1	0.129548
hour_17	0.121468

## Comparison (Non-text features)

### - Logistic Regression



### - Naive Bayes



# Preprocessing and Modeling – Text features (NLP)

- **Preprocessing:**

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

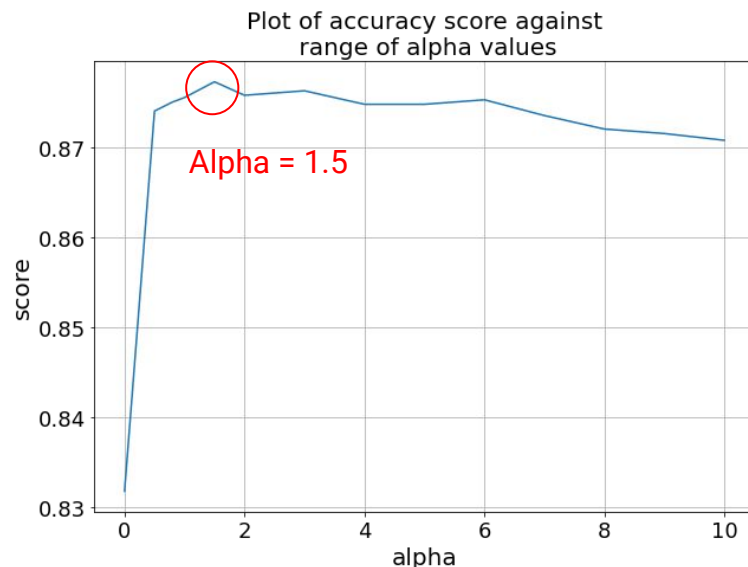
Illegal Immigrants Returning To Mexico For American Jobs	columbus	activist	afghan	afterlife	allegedly	allows	angeles	animatronic
News: Box Office Failure: The \$450 Million 'The Last Jedi' Made On Its Opening Weekend Is Less Than A Third Of What The Cast And Crew Spent On Paper Towels	loan	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.377964
	whenever	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
	senseless	0.447214	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
	halloween	0.000000	0.000000	0.000000	0.383687	0.000000	0.000000	0.000000
News: Accessibility FTW! Tic Tac Is Making Its Breath Mints 500 Times Larger For The Visually Impaired	gunned	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
	employee	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
	settling	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Middle-Aged Woman Angrily Demanding Price Check Was Once Carefree Youth, Onlookers Speculate	play	0.000000	0.000000	0.000000	0.000000	0.000000	0.395751	0.000000
	antigovernment							

# Preprocessing and Modeling – Text features (NLP)

- **Modeling:**
  - Logistic Regression
  - K-Nearest Neighbors
  - Naive Bayes
- Hyperparameter tuning thru GridSearchCV
  - Optimized alpha value of 1.5

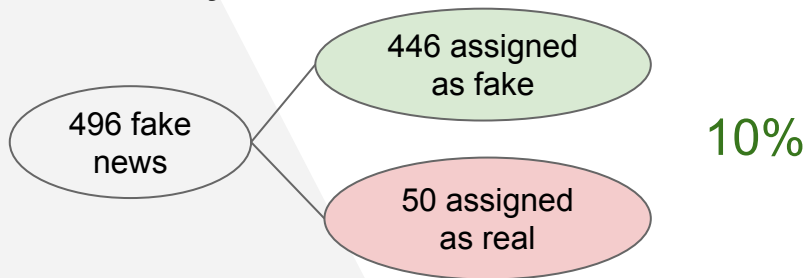
## Best model: Multinomial Naive Bayes

	Train Accuracy	Test Accuracy
Before hyperparameter tuning	0.875	0.829
After hyperparameter tuning	<b>0.877</b>	<b>0.904</b>

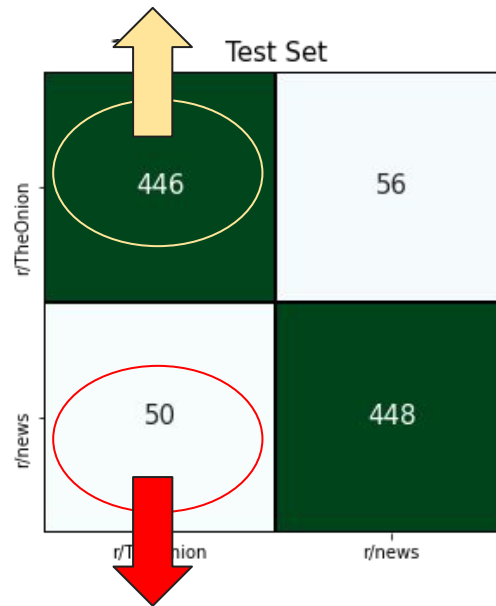
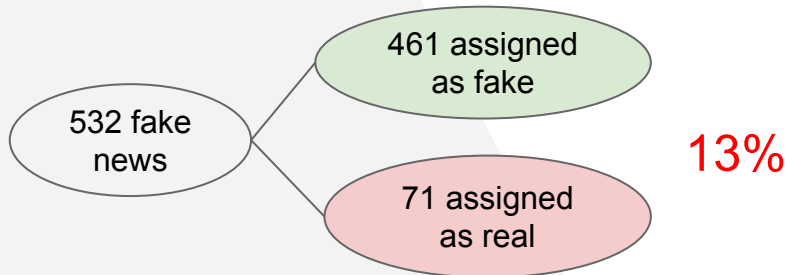


## Comparison (Text features)

- **Naive Bayes**



- **Logistic Regression**



# 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	<b>0.907</b>	<b>0.706</b>	<b>0.743</b>
Precision	<b>0.919</b>	<b>0.681</b>	<b>0.722</b>
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	<b>0.888</b>	<b>0.894</b>	<b>0.78</b>
Precision	<b>0.905</b>	<b>0.889</b>	<b>0.713</b>
Recall	0.865	0.90	0.933
F1-Score	0.885	0.894	0.809

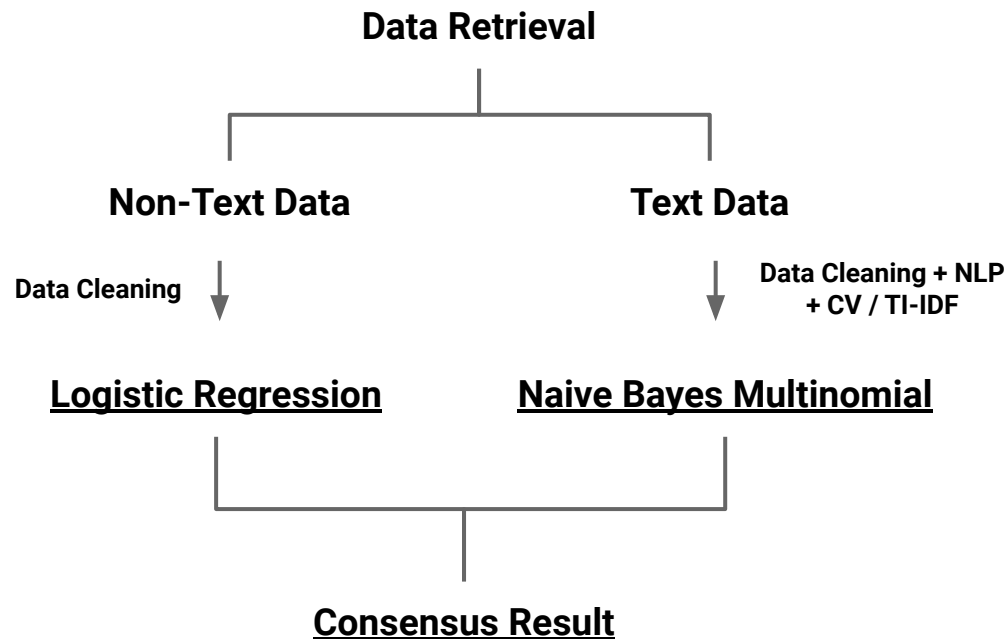
# Evaluation: Combining both models

## 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%





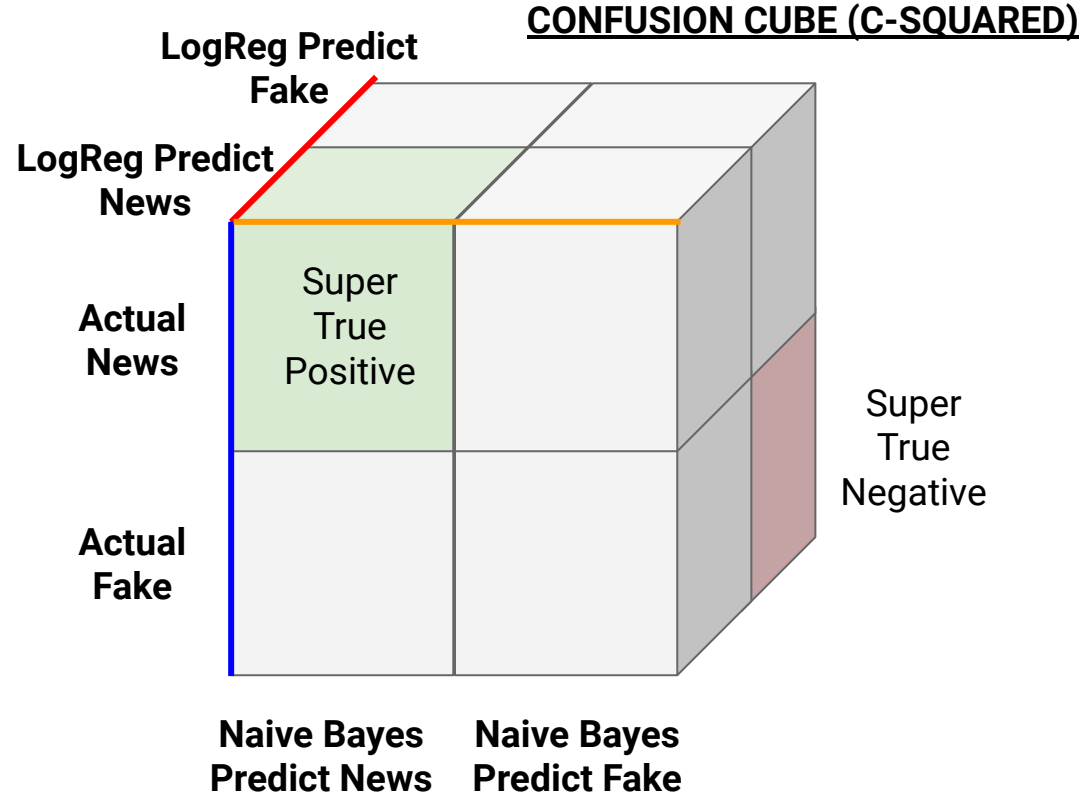
# Extra explanation of consensus result

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



# 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**

**Sri Lanka replaces  
health minister said  
to promote shaman's  
herbal syrup as a  
Covid remedy**

**REAL NEWS**

**FAKE 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	<b>0.907</b>	<b>0.706</b>	<b>0.743</b>
Precision	<b>0.919</b>	<b>0.681</b>	<b>0.722</b>
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	<b>0.888</b>	<b>0.894</b>
Precision	<b>0.905</b>	<b>0.889</b>
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

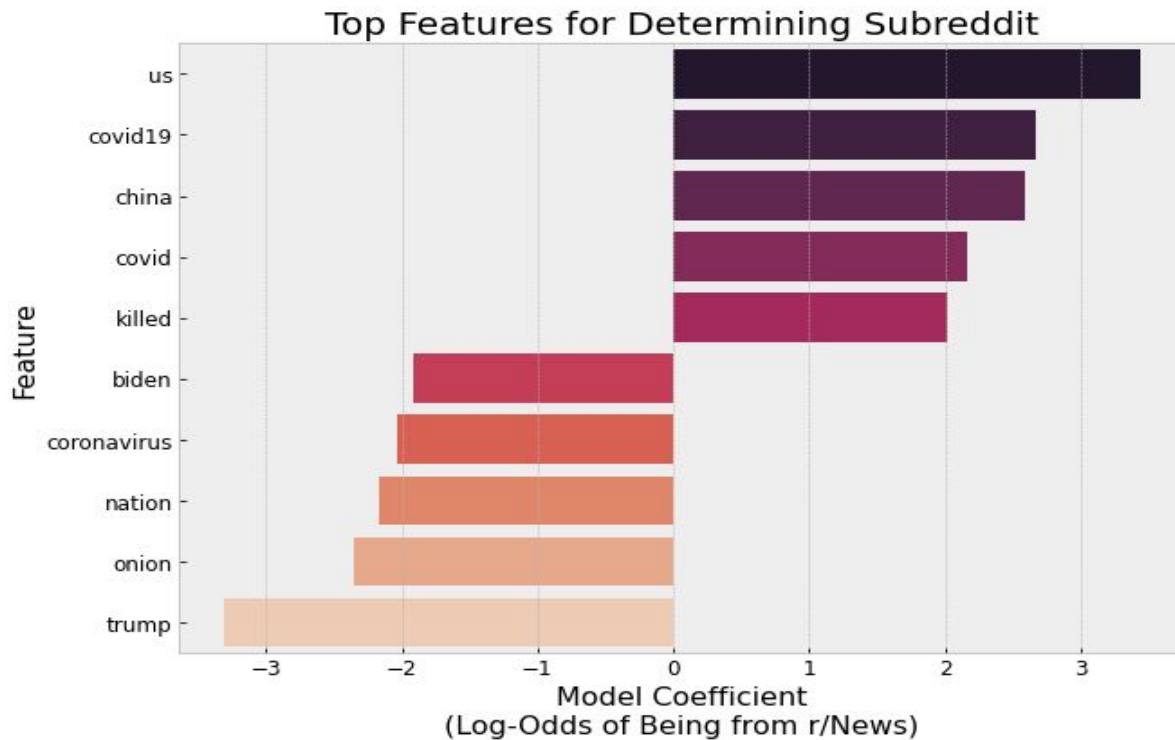
It was interesting to note that Logistic Regression on non-text features had better scores overall when predicting whether a title was legit or not .

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



## News


r/news


### About Community

The place for news articles about current events in the United States and the rest of the world. Discuss it all here.

23.8m	7.7k
Members	Online

---

 Created Jan 25, 2008



## Best of The Onion

r/TheOnion

### About Community

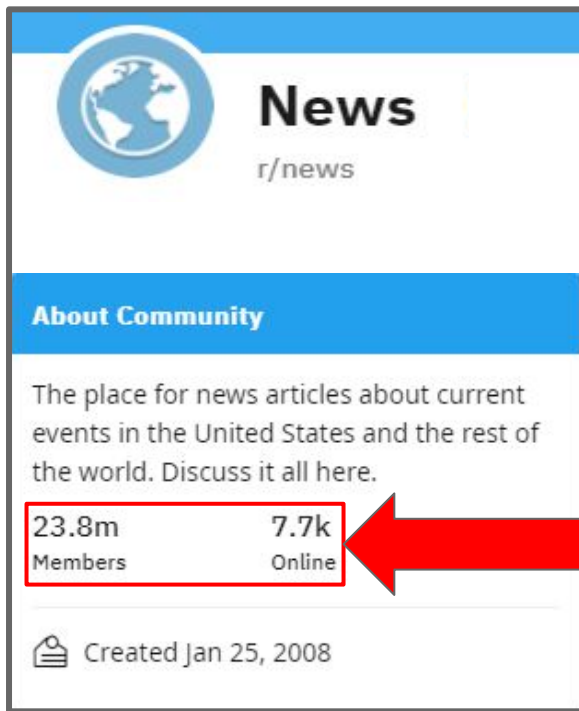
Articles from The Onion. This is not /r/nottheonion. Only links to the Onion are allowed here.

164k	36
Members	Online

---

 Created Mar 23, 2008

# Limitations



The screenshot shows the header of the r/news subreddit with a globe icon and the title "News" above "r/news". Below this is a blue "About Community" section. The description states it is a place for news articles about current events. At the bottom, a red box highlights the statistics: 23.8m Members and 7.7k Online. Below the box, it says "Created Jan 25, 2008".

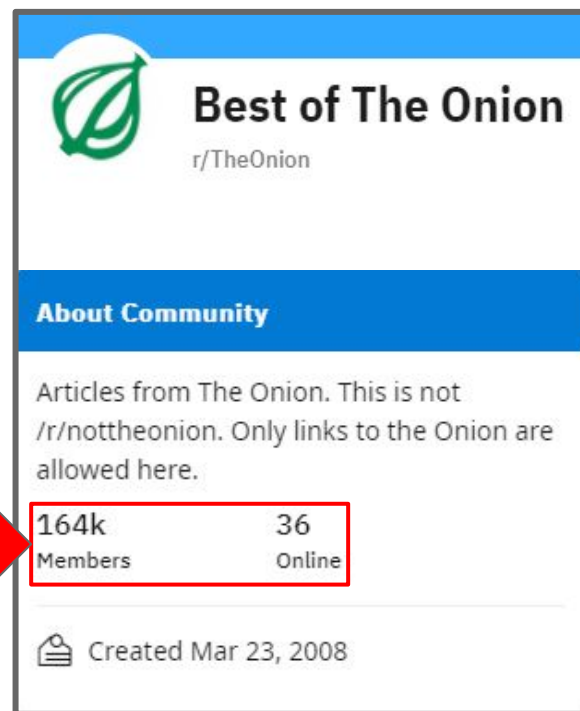
**News**  
r/news

**About Community**

The place for news articles about current events in the United States and the rest of the world. Discuss it all here.

23.8m Members      7.7k Online

Created Jan 25, 2008



The screenshot shows the header of the r/TheOnion subreddit with a green onion leaf icon and the title "Best of The Onion" above "r/TheOnion". Below this is a blue "About Community" section. The description states it contains articles from The Onion and only allows links to the Onion. At the bottom, a red box highlights the statistics: 164k Members and 36 Online. Below the box, it says "Created Mar 23, 2008".

**Best of The Onion**  
r/TheOnion

**About Community**

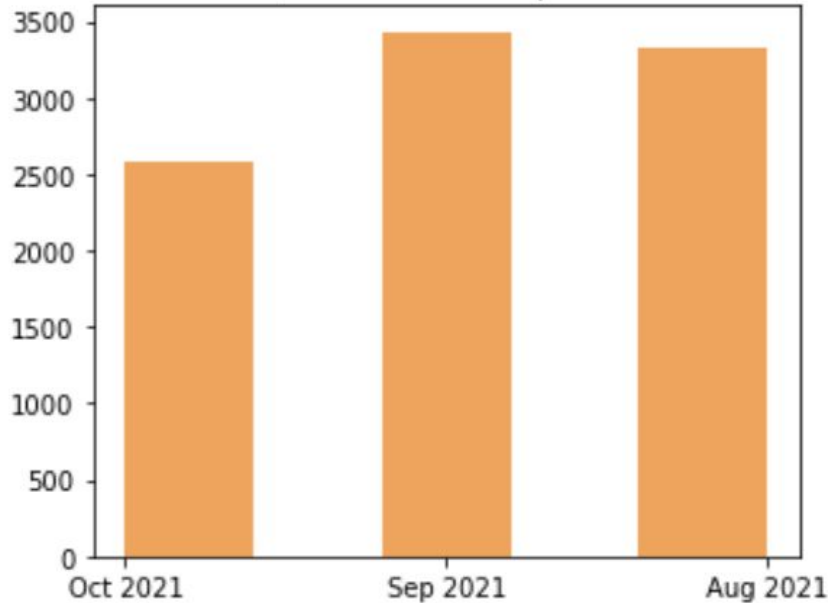
Articles from The Onion. This is not /r/nottheonion. Only links to the Onion are allowed here.

164k Members      36 Online

Created Mar 23, 2008

# Limitations

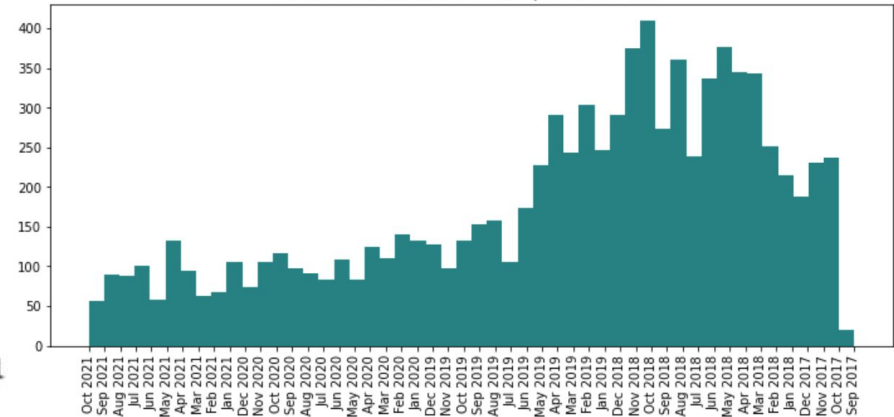
Number of r/news headlines posted over time



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

r/TheOnion headlines spanned back to 2017

Number of r/TheOnion headlines posted over time



[illegible]





A word cloud visualization of the top 100 words from the COVID-19 dataset. The words are arranged in a dense, overlapping manner, with 'death' and 'police' being the largest. A green rectangular box highlights the word 'covid' at the bottom center.







[illegible]

# Future Outlook

- Compare date and headline keywords.

## Future Outlook

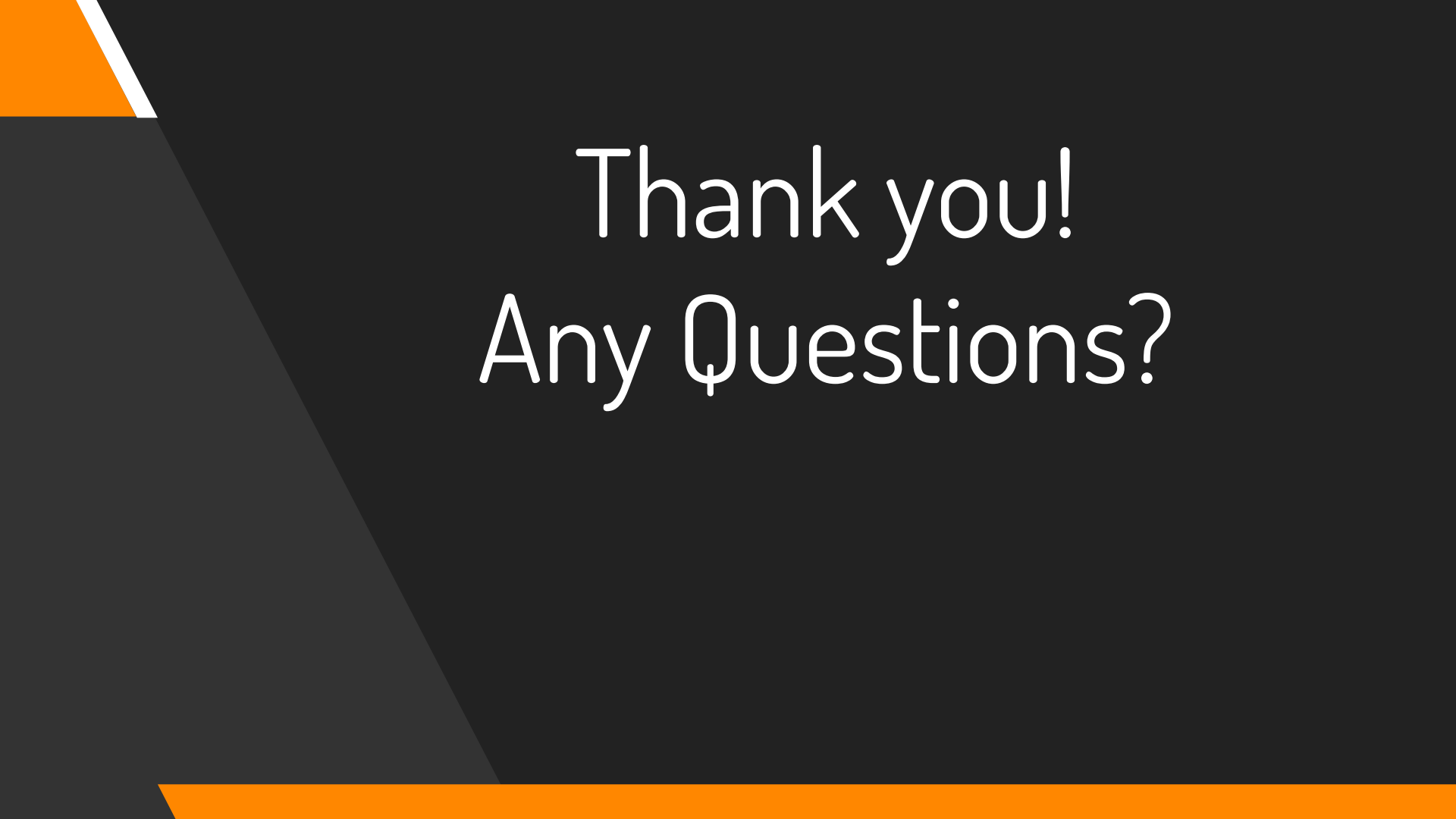
- Compare date and headline keywords.
- Look into other legit news sources to see if the model is able to pull its weight on different headlines from other places.

## 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