



# News headlines on Reddit

Using Data Science to reduce the spread of satire as news





# Table of Contents

**01**

## Introduction

Problem statement

**02**

## Methodology

The data and the process

**03**

## EDA

**04**

## Modelling

**05**

## Conclusion



reddit

Introduction

Methodology

EDA

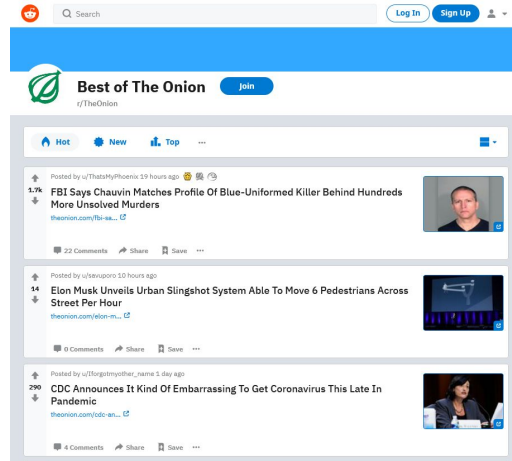
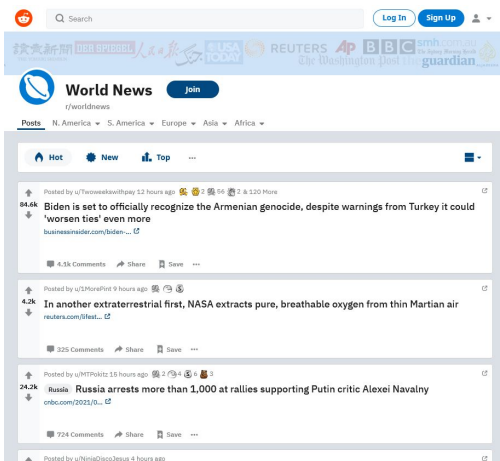
Modelling

Conclusion

(1) | | | preferences | logout

# Overview

- Reddit is a social news aggregation and discussion site.
- Consists of many subreddits
- Popular subreddits include WorldNews and TheOnion



## NEWS IN BRIEF

FTC Rules Businesses Must  
Disclose Whether They Actually  
Cool Or Just Use Minimalist  
Branding



# Problem Statement

- Fake news and misinformation has become a large feature of social media.
- 59% of links shared on social media have never been clicked
- To combat this recurring issue, we have come up with an algorithm to classify satirical and non-satirical news on Reddit
- Our aim is to reduce the spread of misinformation using machine learning algorithms
- Aim to minimise the rate of incorrect satire predictions



# Methodology

- Performed Exploratory Data Analysis
- Transformed words to numbers using CountVectorizer and TfidfVectorizer
  - r/TheOnion stop words added to stopwords list for modeling
- Explore RandomForestClassifier, Multinomial Naive-Bayes, and Logistic Regression models



# Data Collection and Cleaning

- Using Pushshift's Application Programming Interface (API) to collect posts from r/TheOnion and r/worldnews
- Spam posts removed in collection process
- Posts contained only headlines and links
- Heavy focus on headline analysis in data exploration
- Features for character length and word count created from headlines



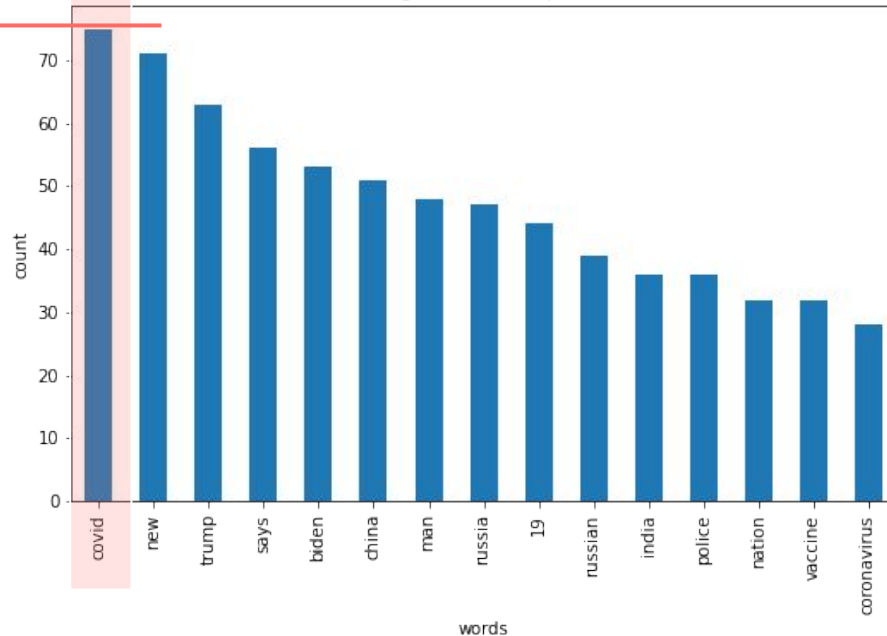
# General observations

- r/worldnews headlines are from a mix of sources
  - Business Insider, CNBC, Reuters, Newsweek
- r/TheOnion headlines exclusively from The Onion
- Top unigrams and bigrams were related to significant events
- r/TheOnion top words had a heavier emphasis on USA current events
- r/worldnews top words had a wider variety with emphasis on COVID-19

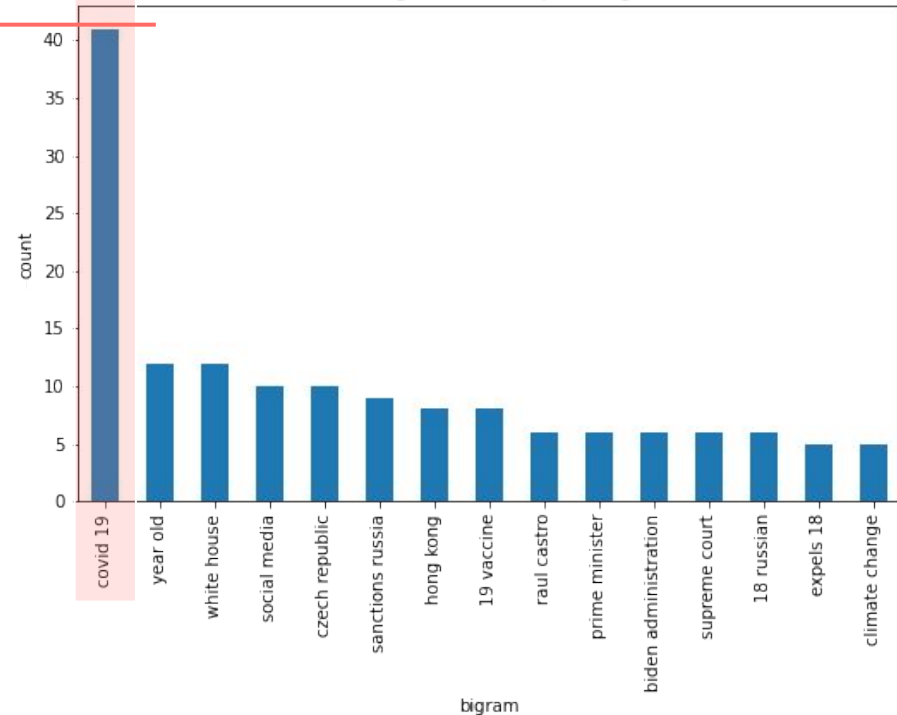


# Countvectorizer had a high bias to 'covid'

Bar chart showing 15 most frequent words in title



Bar chart showing 15 most frequent bigrams in title

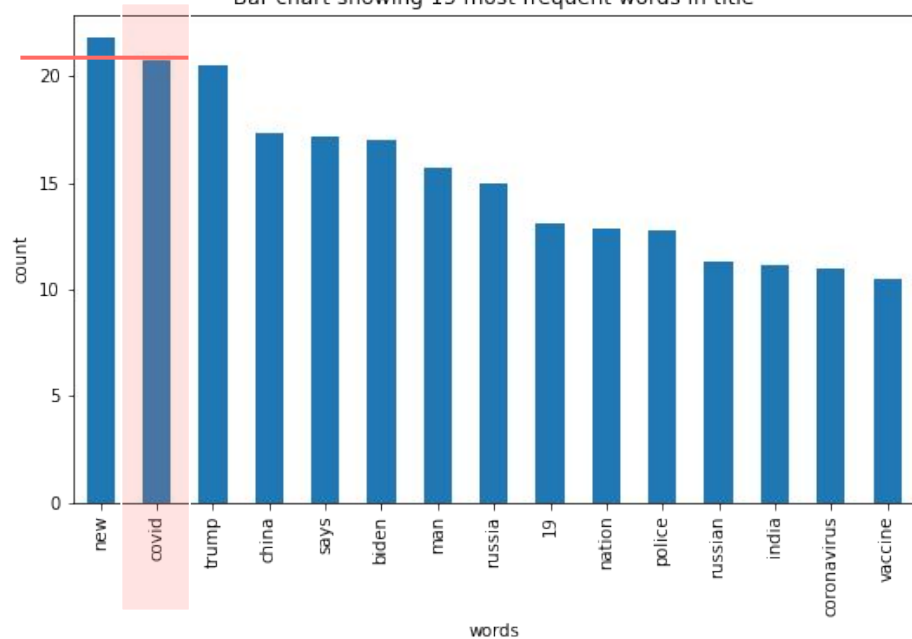




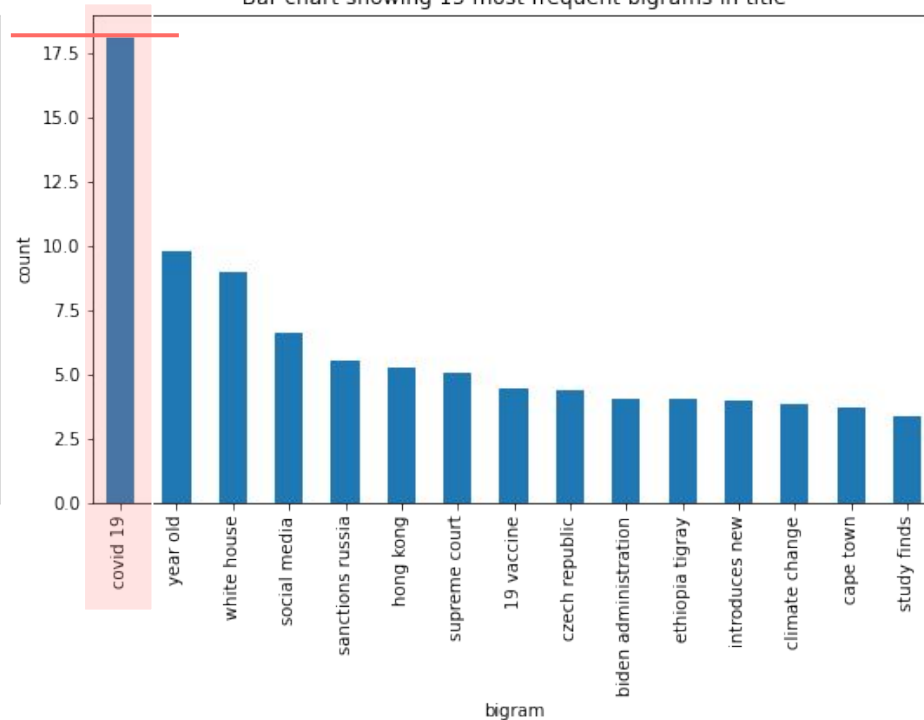


# TF-IDF Vectorizer reduced the bias towards 'covid'

Bar chart showing 15 most frequent words in title



Bar chart showing 15 most frequent bigrams in title





# Modelling Workflow

Cleaned Dataset

Count Vectorizer

Transformers

TF-IDF Vectorizer

Logistic Regression

Multinomial  
Naive Bayes

Random Forests

Estimators

Logistic Regression

Multinomial  
Naive Bayes

Random Forests



# Model Evaluation Metrics

Word & Character Count  
taken into account

Classifier: News = 1  
Baseline Score: 0.48

Cvec log params:  
Stopwords  
Max\_features 800

Cvec log gs params:  
Stop words  
Max\_features 1200

	log	log2	log_gs	log_gs2	tvec_log	tvec_log_gs	nb	tvec_nb	nb_gs	tvec_nb_gs	rf	tvec_rf
<b>BestModelScore</b>	0.860	0.864	0.868	0.871	0.855	0.871	0.862	0.857	0.868	0.865	0.841	0.858
<b>TrainScore</b>	0.924	0.923	0.945	0.948	0.908	0.927	0.892	0.895	0.909	0.914	0.990	0.990
<b>TestScore</b>	0.857	0.860	0.859	0.859	0.853	0.852	0.845	0.845	0.858	0.864	0.836	0.859
<b>TrueNeg</b>	426.000	429.000	427.000	429.000	418.000	421.000	400.000	397.000	400.000	399.000	388.000	420.000
<b>FalsePos</b>	44.000	41.000	43.000	41.000	52.000	49.000	70.000	73.000	70.000	71.000	82.000	50.000
<b>FalseNeg</b>	87.000	87.000	86.000	88.000	83.000	87.000	72.000	69.000	60.000	54.000	68.000	79.000
<b>TruePos</b>	359.000	359.000	360.000	358.000	363.000	359.000	374.000	377.000	386.000	392.000	378.000	367.000
<b>Accuracy</b>	0.857	0.860	0.859	0.859	0.853	0.852	0.845	0.845	0.858	0.864	0.836	0.859
<b>Sensitivity</b>	0.805	0.805	0.807	0.803	0.814	0.805	0.839	0.845	0.865	0.879	0.848	0.823
<b>Specificity</b>	0.906	0.913	0.909	0.913	0.889	0.896	0.851	0.845	0.851	0.849	0.826	0.894
<b>Precision</b>	0.891	0.898	0.893	0.897	0.875	0.880	0.842	0.838	0.846	0.847	0.822	0.880
<b>F1</b>	0.846	0.849	0.848	0.847	0.843	0.841	0.840	0.842	0.856	0.862	0.834	0.851
<b>ROCAUC</b>	0.856	0.859	0.858	0.858	0.852	0.850	0.845	0.845	0.858	0.864	0.837	0.858

- Metrics to optimize are:-
  - Specificity (Reducing False Positives) the closer value is to 1
  - Sensitivity (Reducing False Negatives) the closer value is to 1
  - F1 Score (Balance of Sensitivity and Specificity)
- False Positives refers to posts that are predicted to be news, but is actually satire.
- Important to reduce false positives rather than false negatives as it is more harmful to incorrectly predict satire as news.



# Conclusions and Recommendations

- Logistic Regression model that includes word and character length as features is recommended as the classification model for satirical and non-satirical news on Reddit
- This model can be improved by including the contents of the linked articles to further increase the specificity scoring
  - This will also ensure that the headlines will not be able to mislead and by-pass the classification model if the linked article is actually a satire piece.
- Collecting more data
  - from other satirical news media that is not purely from The Onion to reduce bias
  - More headlines from both satirical and official news sources
- Ideally, we want optimize for both specificity and sensitivity to decrease the both the false positives and false negatives