

Group 🤝

Raymond | Zhe Wei | Ethan | Jia Wen



the ONION®

About Community

r/TheOnion

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

165k

Members

23

Online



Created Mar 23, 2008



NOT
the ONION®

About Community

r/nottheonion

For true stories that are so mind-blowingly ridiculous that you could have sworn they were from The Onion.

20.6m

Readers

8.0k

Online



Created Oct 25, 2008

r/TheOnion or r/nottheonion ??

Posted by u/geekman20 10 hours ago

Girl survives heavy snowstorm by hugging
a dog, Russian state TV reports



Posted by u/dwaxe 10 days ago

'Hero Rat' Who Detected Landmines In Cambodia Dies In Retirement



Architect behind 'dangerous' to

Posted by u/El_Zarco 9 hours ago

Minecraft DDoS Attack Leaves Small European Country

Without Internet

Man Horrified As He Has No Past

Posted by u/noraad 1 day ago

Serial k...

Posted by u/dwaxe 10 days ago

'Hero Rat' Who Detected Landmines In Cambodia Dies In Retirement

PD: Man dies after he's run over by own pickup truck while tree

?

Problem statement

Fake news is a prevalent and harmful problem in our modern society, often misleading the general public on important topics such as healthcare and defense. This can lead to long standing societal issues which are a detriment to nations worldwide.

Goal

Our team aims to develop a model using natural language processing and machine learning models to predict whether an article is from `r/TheOnion` (fake news) or `r/nottheonion` (real news).

Helping government bodies/regular citizens to identify the fake news, thus creating a secure, and more misinformation-resilient society.

Choosing the subreddits



Real News



Fake News



the ONION

In general:

- Worded in similar fashion (particularly the headline)
- Reference real world events and figures
- Difficult to find a consistent source for a wide variety of real news and fake news for our model

In general:

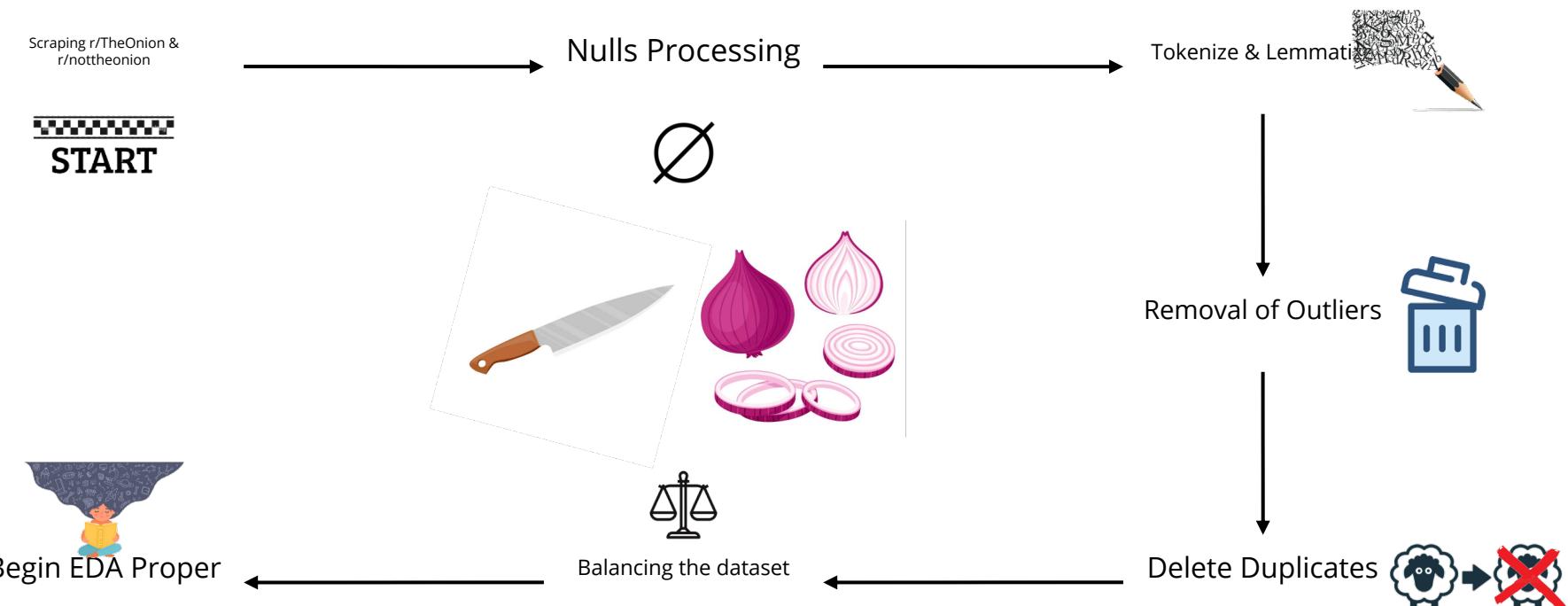
- Worded in similar fashion (particularly the headline)
- Reference real world events and figures.
- Easy to obtain a consistent, and wide variety of real and fake news for our model

Overview

1. Background & Problem Statement
2. Procedures & Methodology :
 - Data collection
 - Data cleaning & EDA
 - Preprocessing
3. Modeling & Evaluation
4. Conclusion & Recommendation
5. Q&A



Processing The Dataset



Data Collection

Use PUSHSHIFT API for the data collection

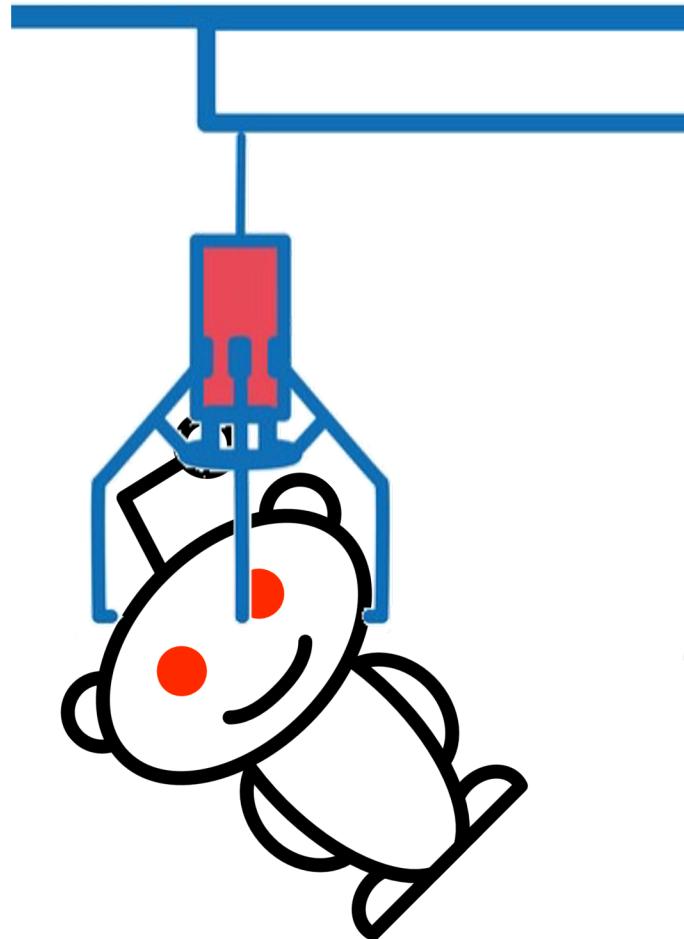
Considerations:

- Latest date **1st January 2022, 0:00 hrs, (GMT+8)**
- Pull at least 5,000 posts from each subreddit
- Delete duplicates as we pull in blocks of 100 posts
- Combine the dataframes as we pull

Finally, we save the 2 sets of data as .csv files.

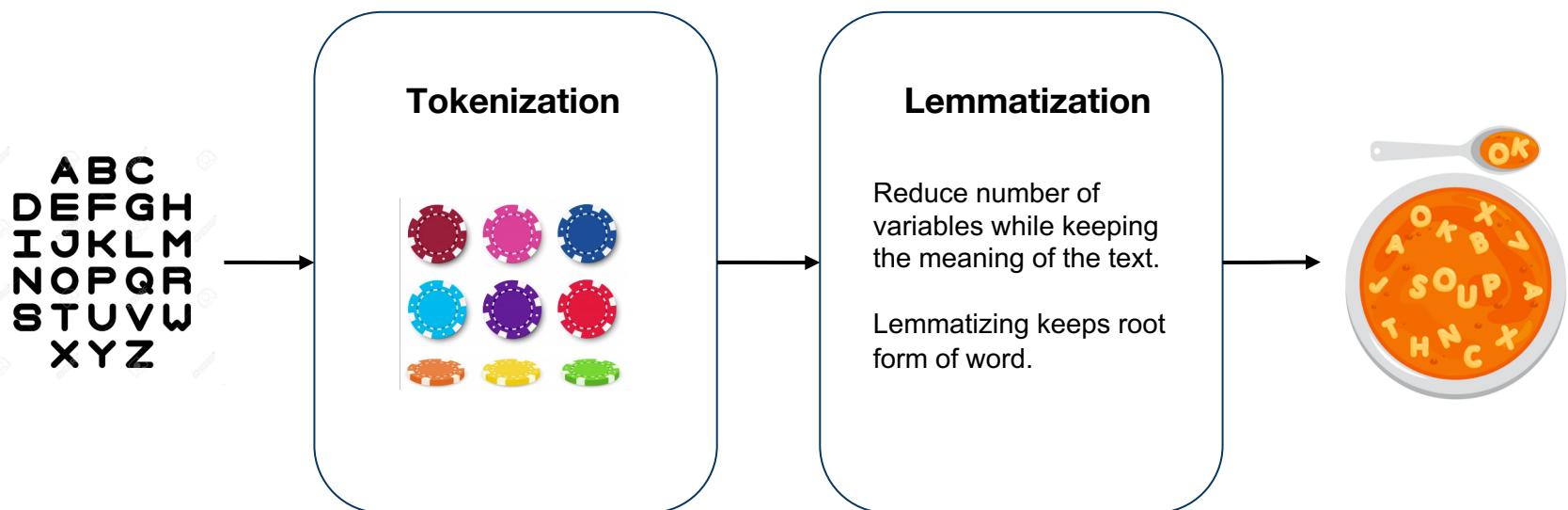
There were no nulls in the dataset for the relevant columns!

[‘ subreddit’, ‘ title’, ‘ created_utc’]



Tokenization & Lemmatization

Look at the Title text only! Title text is the Headline.



Data Cleaning



Single Emoji

Lightroom Single Words

[https://www.yahoo.com/entertainment/sports-
rad...](https://www.yahoo.com/entertainment/sports-rad...) Single URLs

| nothing happened 2 word titles

强烈批评! 塔利班内阁缺少其他极端主义代表

Non-english
characters

The onion

Contains the
word 'onion'



utliers In Title Text!

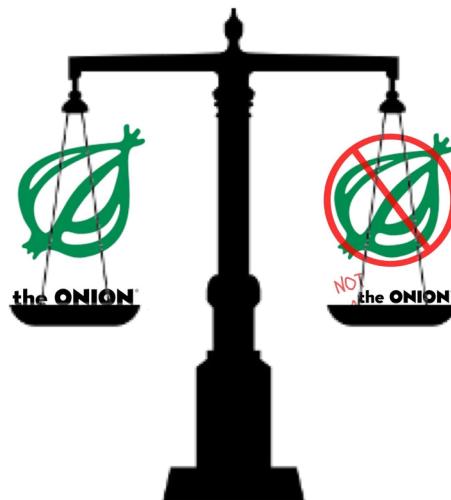
Balancing the Dataset

5,000 posts from r/TheOnion

5,000 posts from r/nottheonion

Filtered by UTC/date -

5,000th post is the earliest

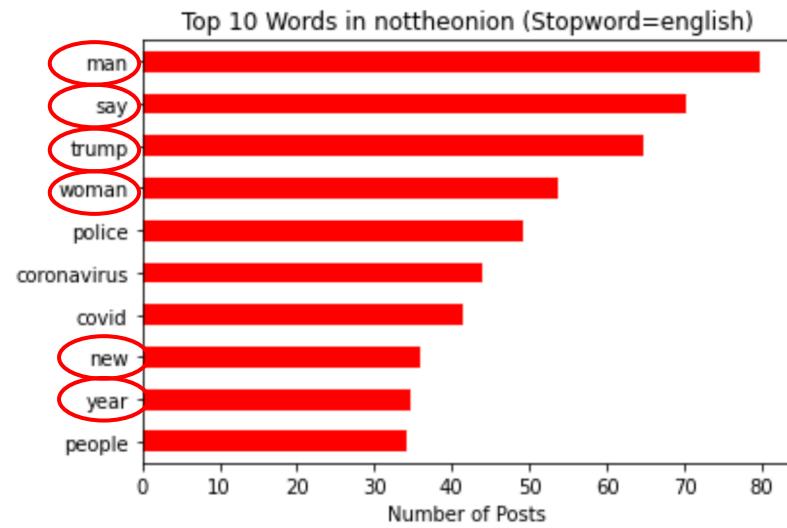
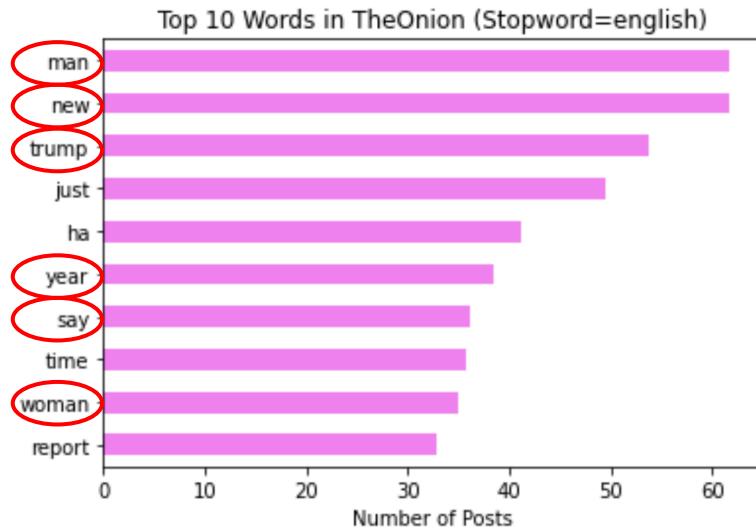


Balanced dataset is important for a classification problem!

It ensures that no one class takes precedence over the other.

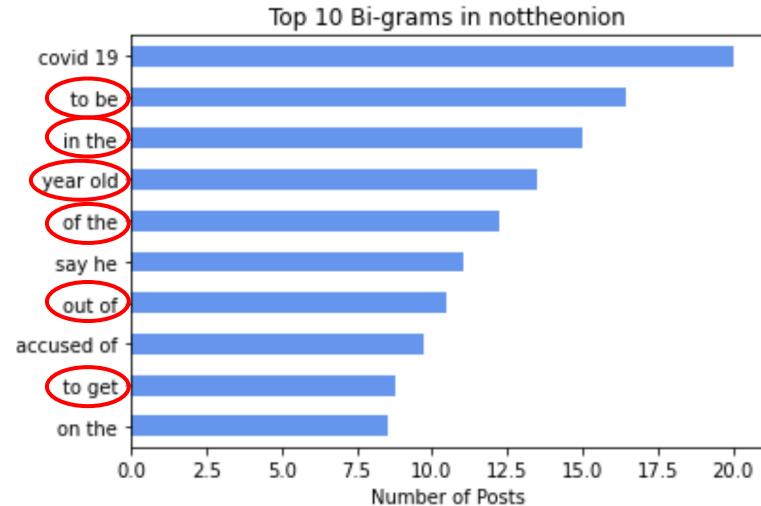
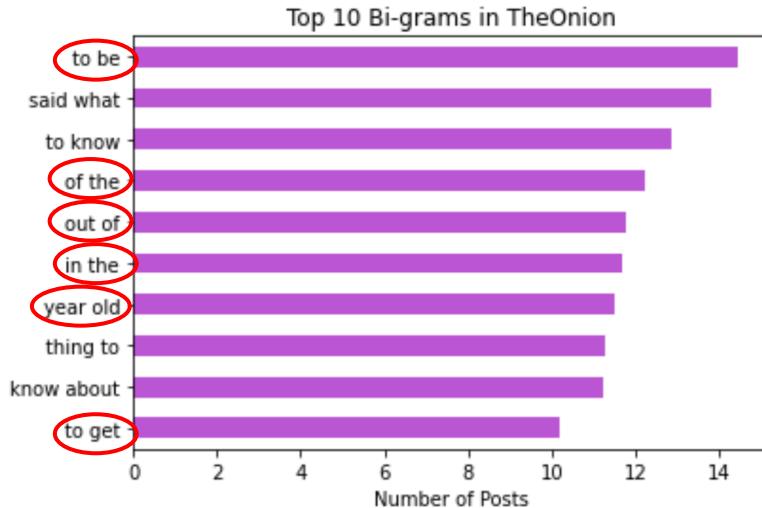
Overall more balanced and fair metrics such as accuracy, recall, precision and f1 score.

EDA - Top Words in Dataset



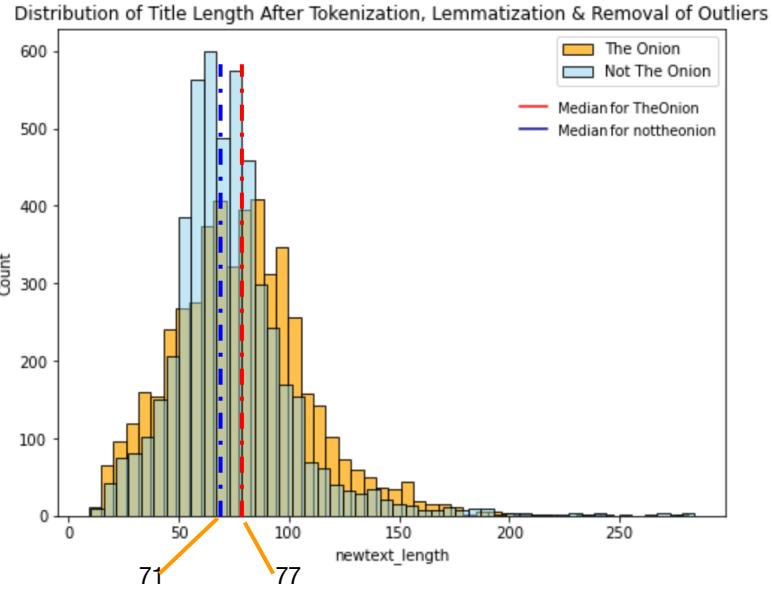
Common words for both r/TheOnion and r/nottheonion are '**new**', '**man**', '**woman**', '**trump**', **year** and '**say**'.

EDA - Top Bi-grams in Dataset



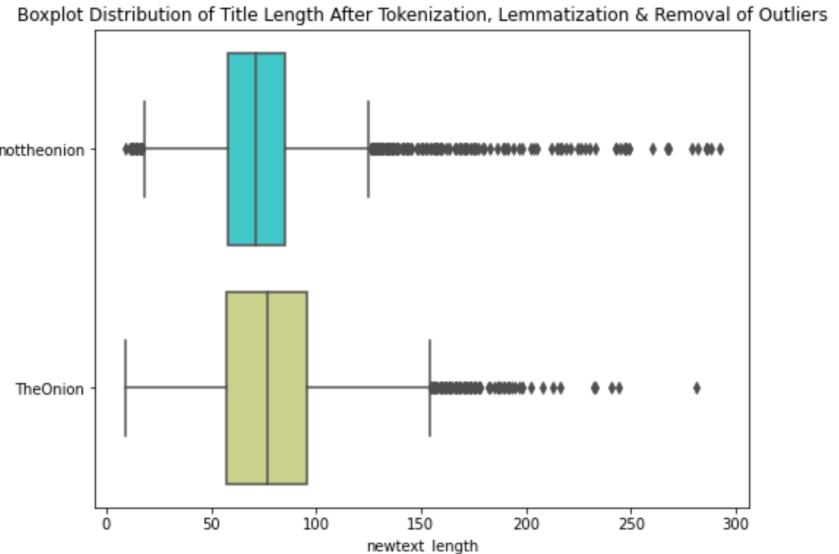
Common bi-grams for both r/TheOnion and r/nottheonion are ‘to be’, ‘in the’, ‘out of’, ‘of the’, ‘to get’ and ‘year old’.

EDA - Title Length

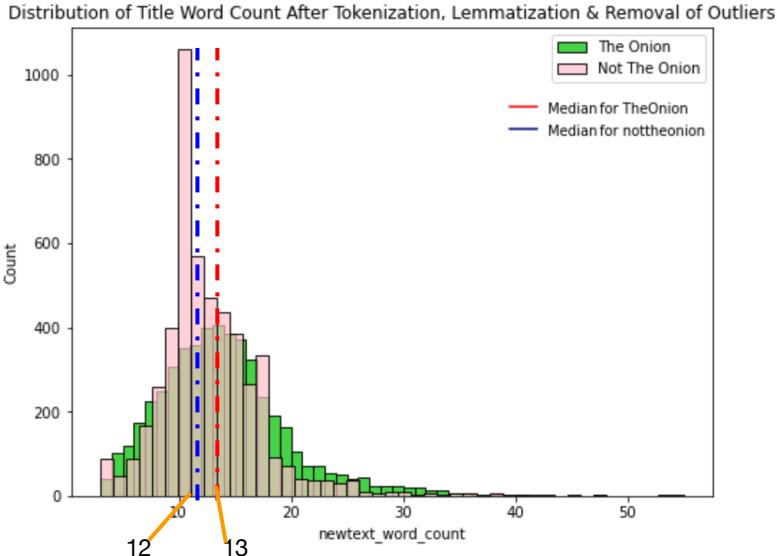


For r/TheOnion, the title length peaks between the 60 - 90 range.

For r/nottheonion, the title length peaks between the 60 - 75 range.

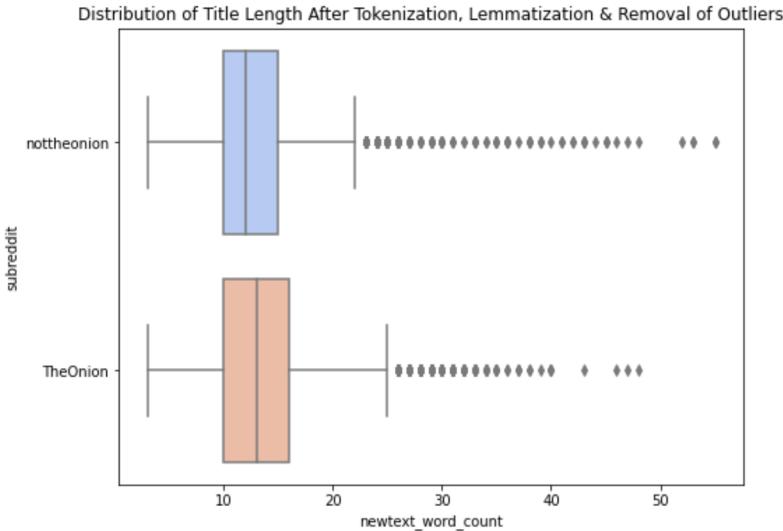


EDA - Word Count



For r/TheOnion, the word count peaks between the 10 - 15 range.

For r/nottheonion, the word count peaks between the 9 - 14 range.

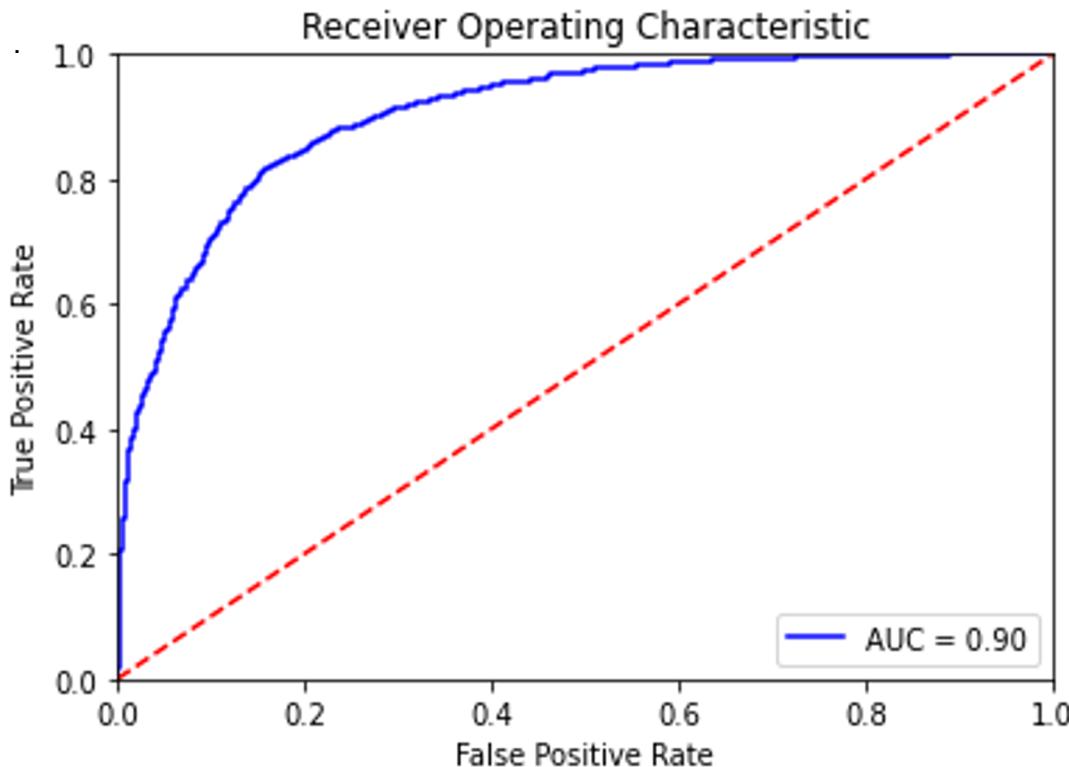


Model

- Passed the clean dataset into a countvectorizer and passed it through various classification model.
- Metric used: F1 score which balances precision and recall.

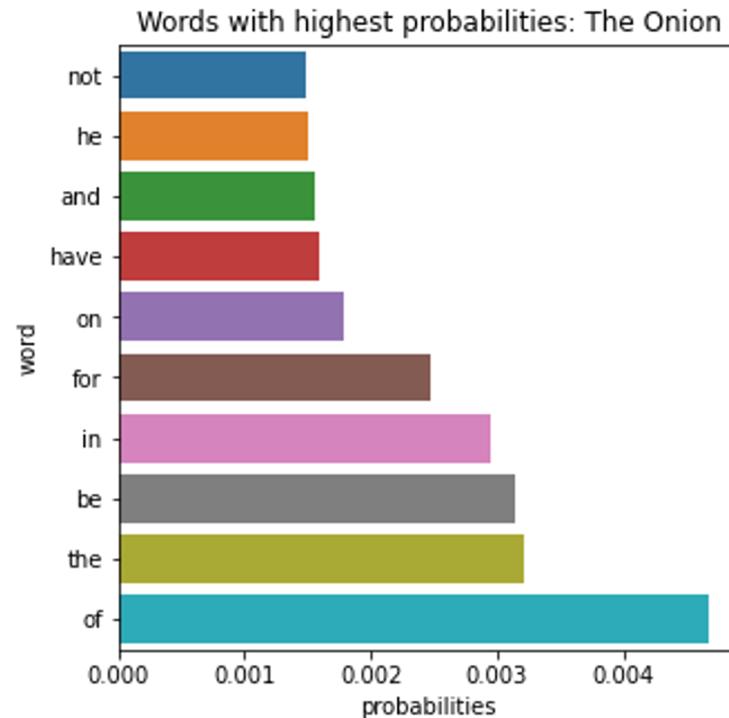
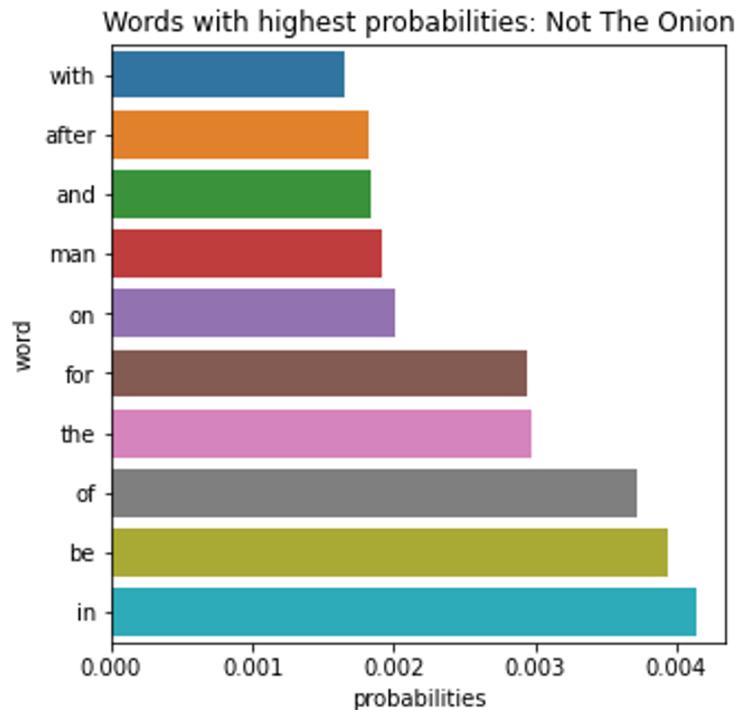
Model - Place holder Table	Performance	Tuned
Baseline	0.5	-
KNN	0.6	-
NB	0.81	0.82
LR	0.80	0.82
Random Forest	0.79	0.82

Model - ROC

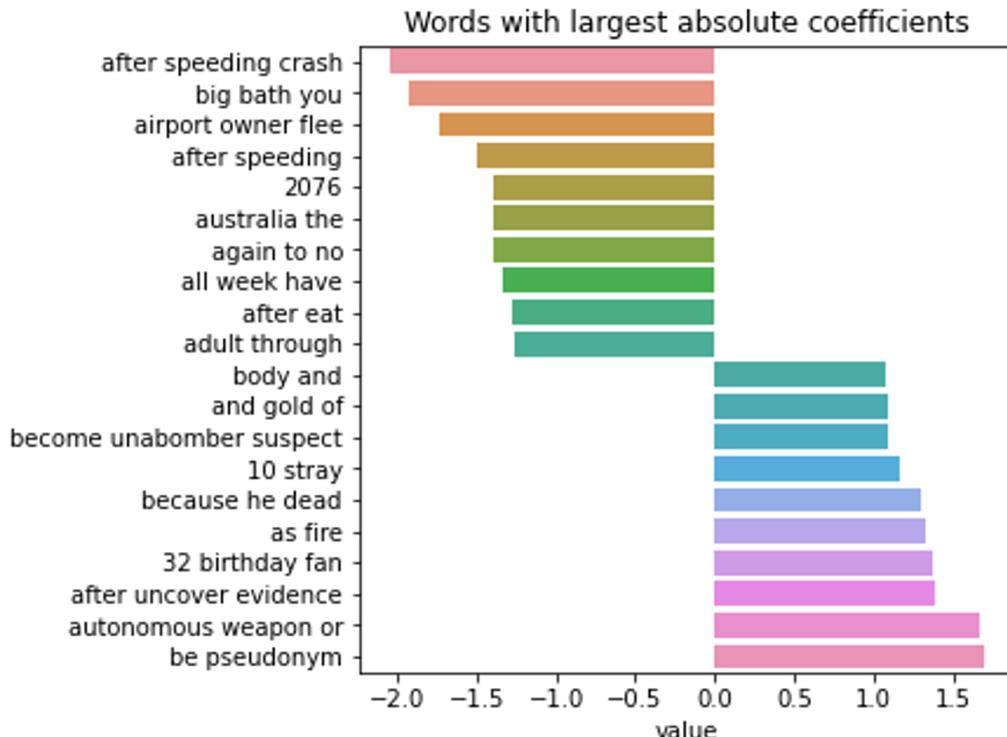


- The ROC plot shows a fairly smooth curve with an AUC of 0.90 the model is a 90% chance that the model can distinguish between a post from the onion and not the onion.
- Optimal classification threshold is ~0.8 where we can maximise True positives will keeping FPR at a manageable level.

Key Findings - Multinomial Naive Bayes

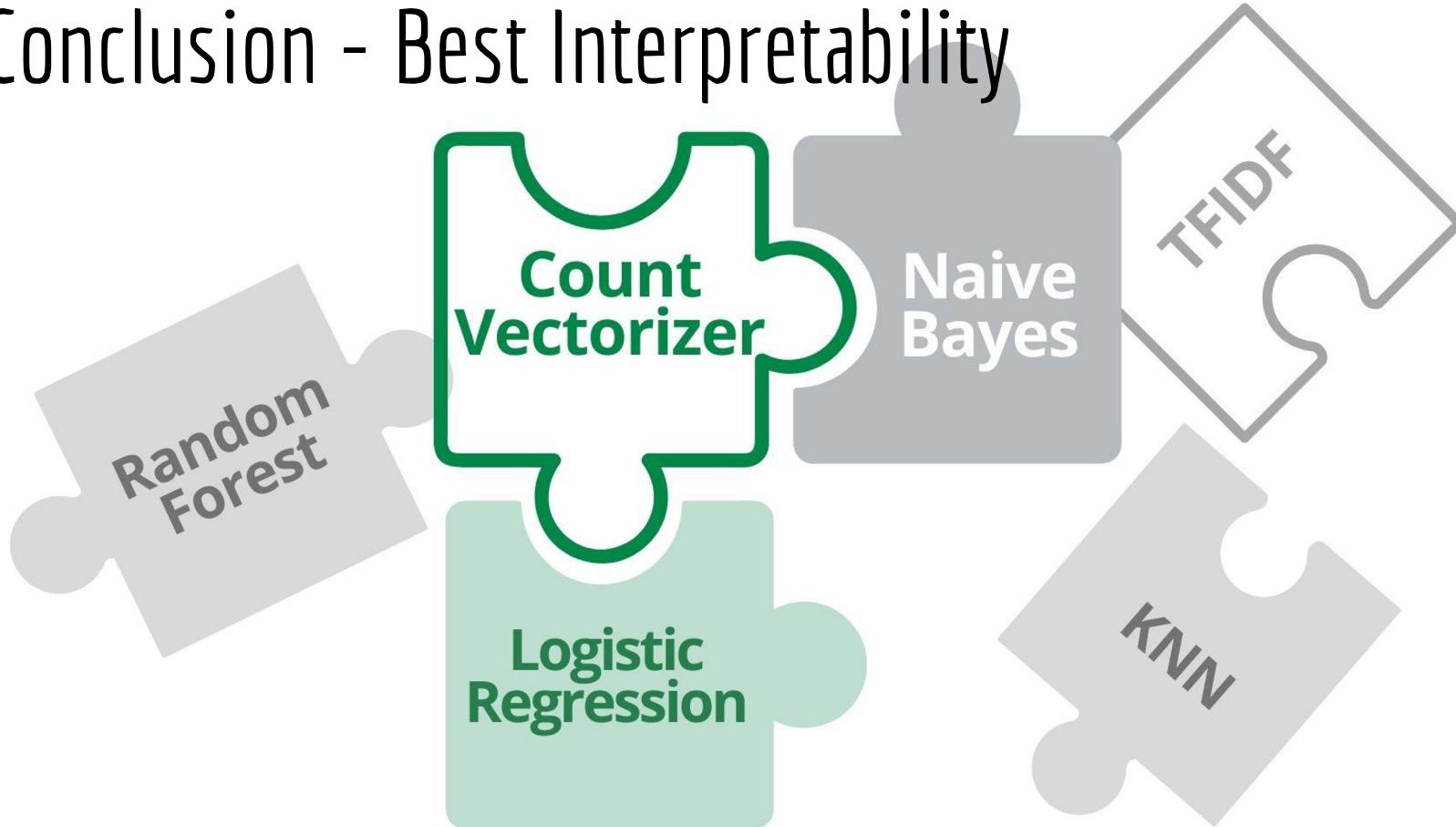


Key Findings - Logistic Regression

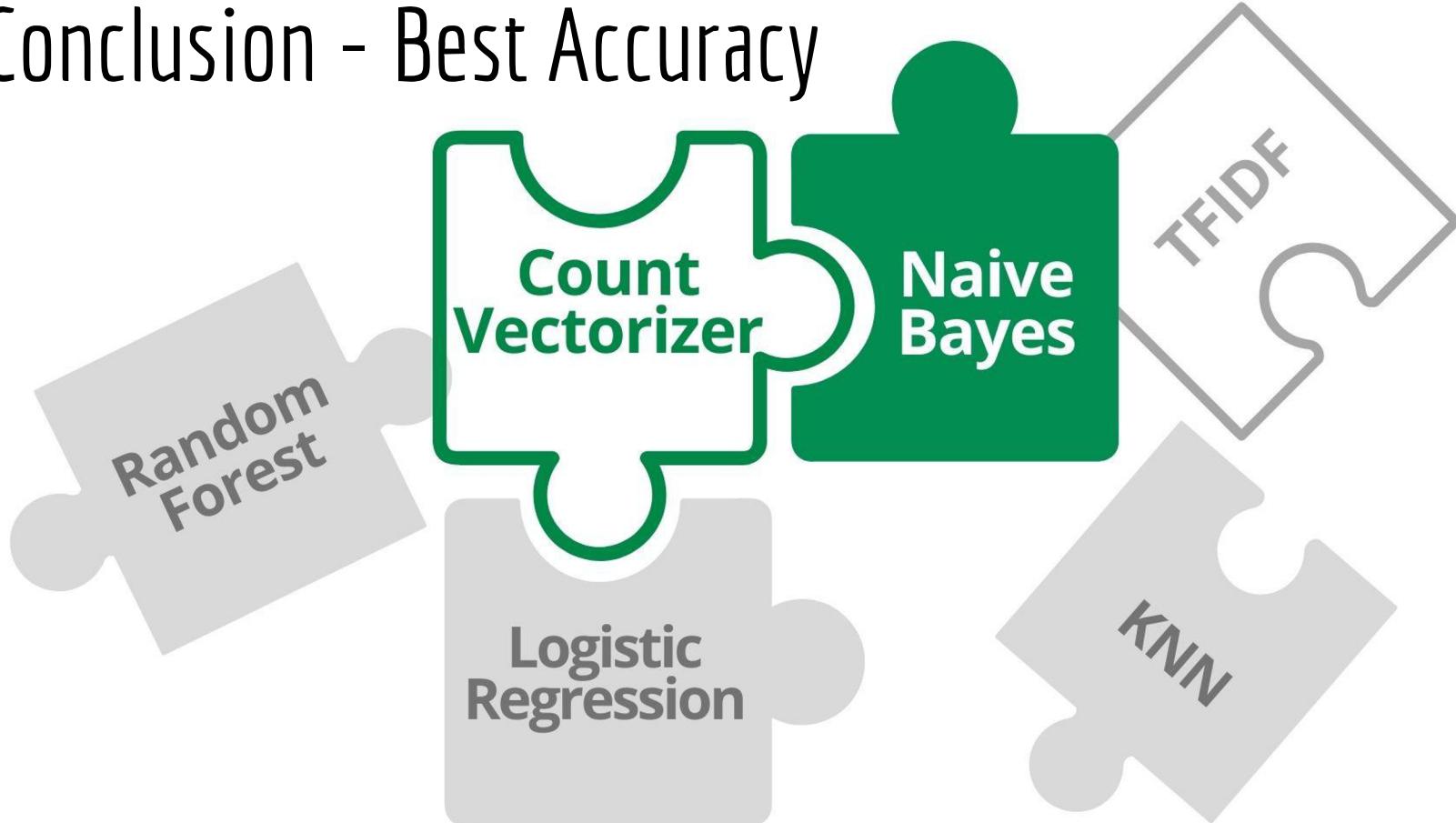


- Compared to LR, by referring to the regression coefficients, the reasons for the models classification are more easily understood.
- Depending on how interpretable we need our model to be we can utilise different logistic regression instead

Conclusion - Best Interpretability



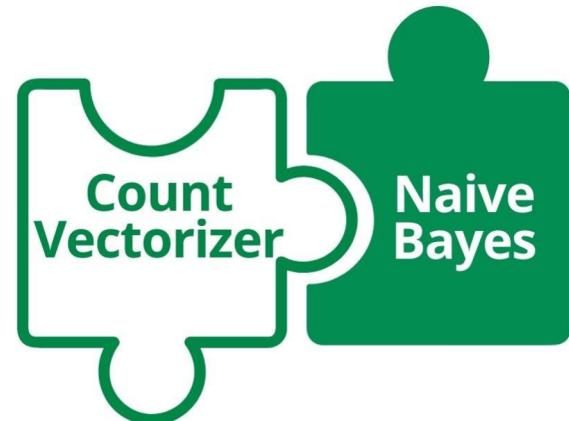
Conclusion - Best Accuracy



Conclusion

CountVectorizer + Naive Bayes Model

- Best F1-Score
 - Least False Positives and False Negatives
- Simple, easy to implement
- Good, accurate text classification prediction



Problems



- Overfitting
 - Train set accuracy (90%) > Test set accuracy (82%)
- Limited to English words
- Might not be suitable for future classifications
 - New acronyms and words
 - Recentness of the terms used
 - Shift in topic of concern/discussion

Possible Enhancements



- Non-text titles
 - images, videos
- Other post features
 - subtext, comments, upvotes
- Post authors
 - analyse authors' posting history
- Content-based analysis
- Other NLP techniques
 - BERT

Thank You!

