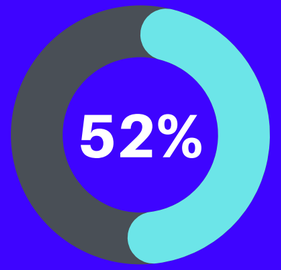
FactGuard

An ML and NLP based fake news classifier

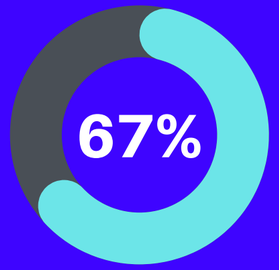
Introduction

- Fake news is information that is explicitly incorrect & misleading. It is used to harm a person or entity’s reputation or to profit from advertising income.
- With the rise of social media and the sheer volume of people who use the internet, there is no shortage of online objects vying for our attention.
- The ease with which people may republish content & the lucrative nature of publishing on social media has allowed fake news to flourish.
- These days, fake news is considered a major threat to countries worldwide. Hence, it is of utmost importance we find a way to inhibit its spread if not stop it entirely.

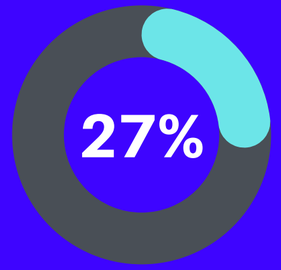
Some Fake News Statistics



regularly encounter fake news online

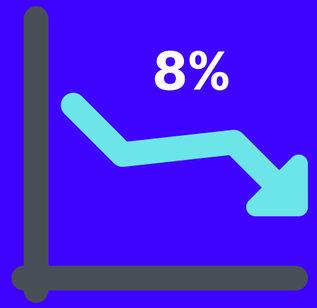


come across false information on social media



trust the mainstream news media

fake news gets over 1.8 billion engagements per month on social media networks



worldwide media trust has dropped by 8% b/w 2021 and 2022

WhatsApp deleted over 2 million accounts sharing misinformation in India

*Sources: Edelman, Statista, Ipos, German Marshall Fund

Real world example

IT'S OVER: Hillary's ISIS Email Just Leaked & It's Worse Than Anyone Could Have Imagined...

POSTED BY FRIENDSOFSYRIA IN WAR CRIMES ≈ 351 COMMENTS

– Hillary Clinton, Friend of the Syria people? Like the USA is friends of the people of Iraq, Afghanistan, Pakistan, Libya, Somalia, Yemen...?

Today Wikileaks released what is, by far, the most devastating leak of the entire campaign. This makes Trump's dirty talk video looks like an episode of Barney and Friends.

Even though when Trump called Hillary the 'founder' of ISIS he was telling the truth and 100% accurate, the media has never stopped ripping him apart over it.

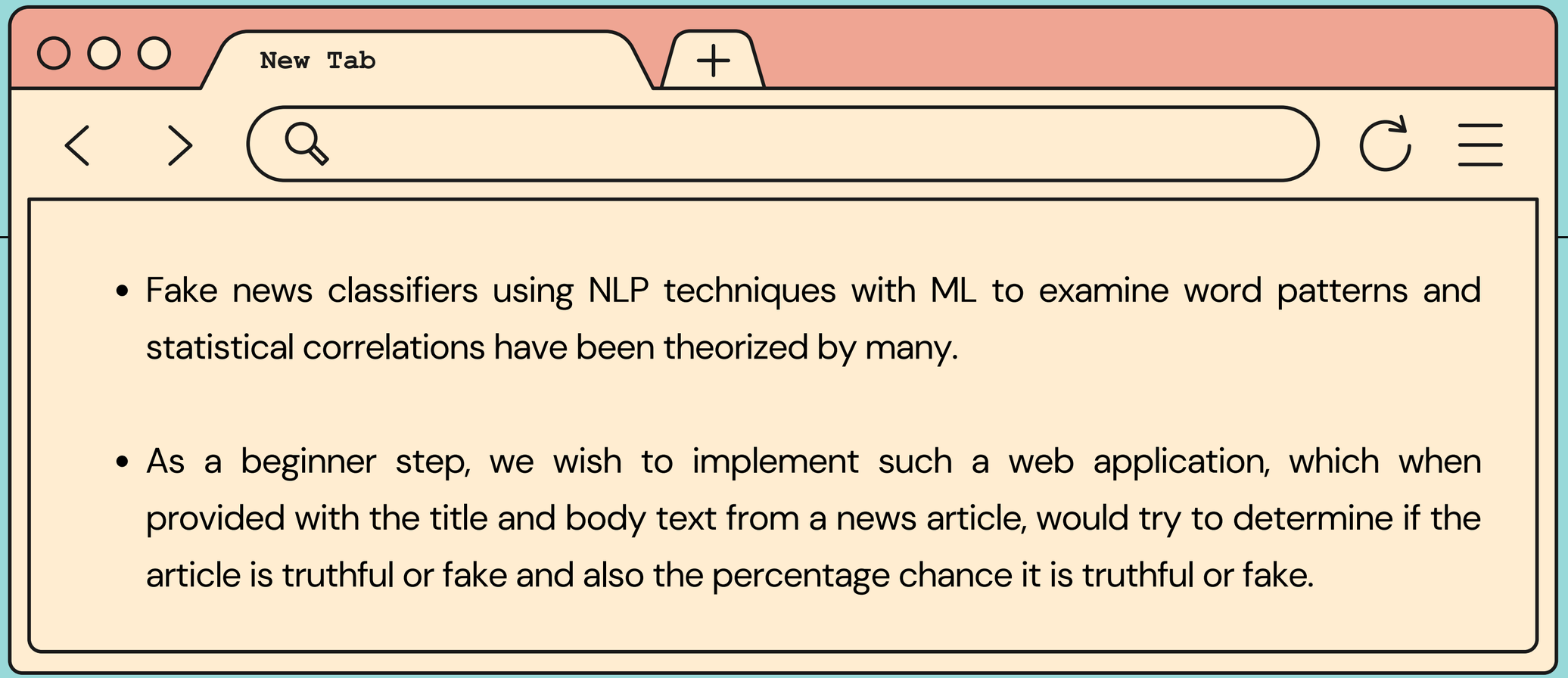
Today the media is forced to eat their hats because the newest batch of leaked emails show Hillary, in her own words, admitting to doing just that, funding and running ISIS.

John Podesta, Hillary's campaign chair, who was also a counselor to President Obama at the time, was the recipient of the 2014 email which was released today.

Assange promised his latest batch of leaks would lead to the indictment of Hillary, and it looks like he was not kidding. The email proves Hillary knew and was complicit in the funding and arming of ISIS by our 'allies' Saudi Arabia and Qatar!

- A fake news story with a sensationalized title
- Originally published on Ending the Fed
- Got ~754,000 engagements in the final three months of the 2016 U.S. presidential campaign
- It is in the top 3 most performing fake election news stories on Facebook.

Objective

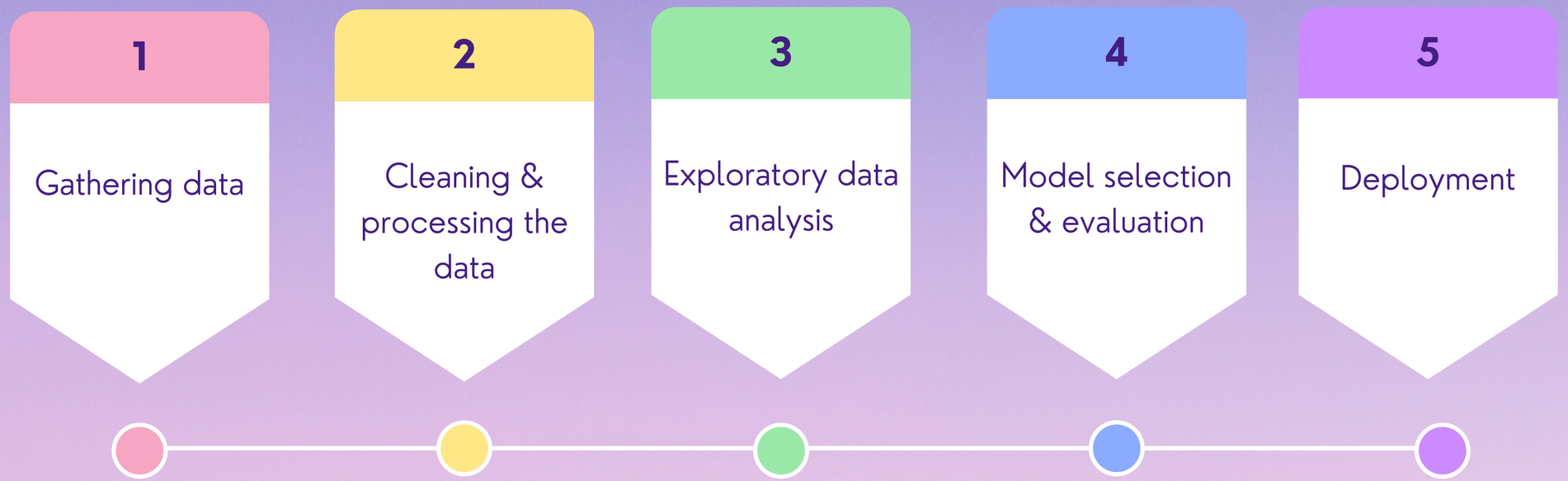


- Fake news classifiers using NLP techniques with ML to examine word patterns and statistical correlations have been theorized by many.

- As a beginner step, we wish to implement such a web application, which when provided with the title and body text from a news article, would try to determine if the article is truthful or fake and also the percentage chance it is truthful or fake.

Methodology

We have decided on the below given methodology for working on the project (Note that it is a high level view and doesn't go into specifics):



Involves manual scraping of data using scripts from various sources or just using existing datasets.

Crucial step which involves techniques like tokenization, stop-words removal and so on. Data collected from different sources have their parameters normalized in this step.

The processed data is explored to gain an understanding of any patterns, trends or anomalies that can be used to aid in modelling.

Various classification models are run over the data to decide which model is the best fit. Performance metrics are then optimized if feasible.

The trained model is deployed online so as to turn it into a GUI with for easy user access.

The Process

Three fundamental steps was undertaken while implementing this application – the first one being gathering the dataset, second being exploratory data analysis (EDA) and finally, modelling & evaluation.

Gathering Data

- 34,894 stories, which includes news from both fake and truthful sources. Multiple datasets (primary & supplemental) used to improve accuracy.
- Clean & filter data using various parameters for exploratory analysis followed by text pre-processing for modelling.

EDA

- Check the distribution of labels.
- Explore various paramters like length of title, having capitals in title & text, news organizations mentioned and frequently occuring words in each case.

Model & Evaluate

- Compare and contrast the model iterations.
- Tune the various parameters.
- Gather insights from the model's performance to see make improvements.

Primary Dataset

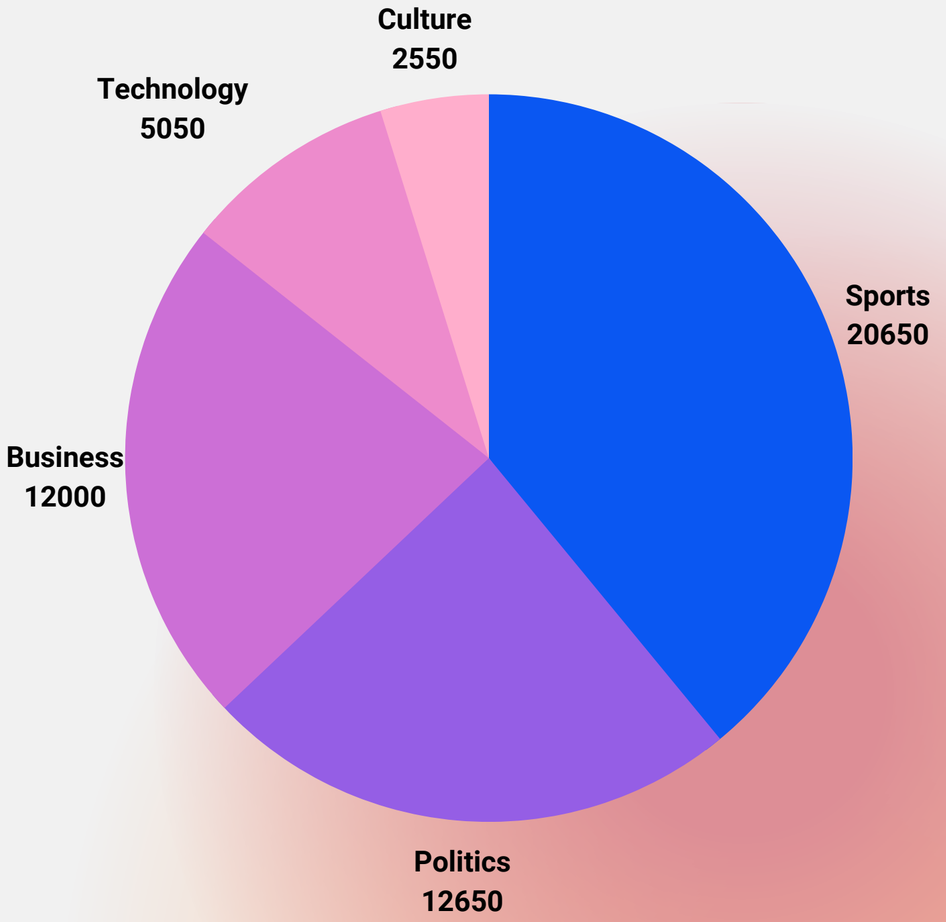
- The data was originally collected by the University of Victoria ISOT Research Lab from real-world sources.
- Truthful articles were obtained by crawling articles from reuters.com (News website).
- Fake news articles were collected from different sources, usually unreliable websites that were flagged by Politifact (a fact-checking organization in the USA) and Wikipedia.
- The dataset contains different types of articles on different topics, however, the majority of articles focus on political and World news topics, between 2015 - 2018.

News	Size (Number of Articles)	Subjects
Real news	21,417	World News: 10,145 Politics: 11,272
Fake news	23,481	Government News: 1,570 Middle East: 778 US News: 783 Left News: 4,459 Politics: 6,841 News: 9,050

*Note: citation, along with kaggle link will be given at end.

Supplemental Dataset

- We use a supplemental dataset because our primary dataset only contains real news stories from a single source (Reuters), which might lead to model overfitting despite our efforts to not do so.
- The supplemental dataset was gathered by a user named Sameed Hayat on Kaggle, from The Guardian (news website).
- It contains over 52,000 real news articles on several topics, but the only ones used here are from the politics section (contains approx. 12,650 articles).



Dataset composition

*Note: kaggle link to dataset will be given at end.

Cleaning & filtering the primary dataset

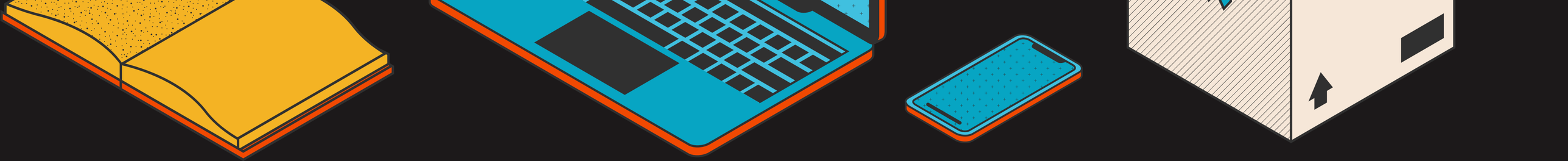
Before we can perform exploratory data analysis on the dataset, we have to clean and filter it. In case of our primary dataset, we performed the following operations to clean & filter it:

1. Check for missing data: First, we checked if there were any NULL values in any of columns.

2. Check for placeholder values & duplicates: Next, we checked if there were any placeholder values being used or any articles were being repeated.

3. Check date validity: Following that, we checked the date fields for a valid format.

4. Check date range: Finally, we checked the date range for news stories and decided whether to keep/delete them according to results.



Results

The following findings were made from cleaning & filtering the data:

Missing data

Fortunately, there were no missing data ("NULL" values) in this case.

Date validity

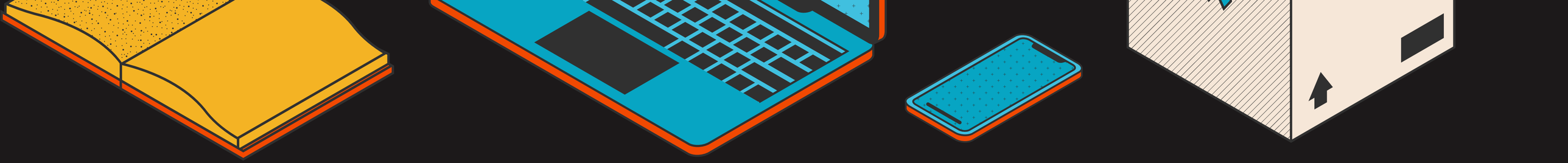
It was found that not all date values are dates. Some contained links and other had junk data.

Fix: All instances with text in date were considered invalid and deleted.

Date range

The description of the data indicated that it was between 2016 and 2017. We checked data that was outside this range and found that all were labelled fake.

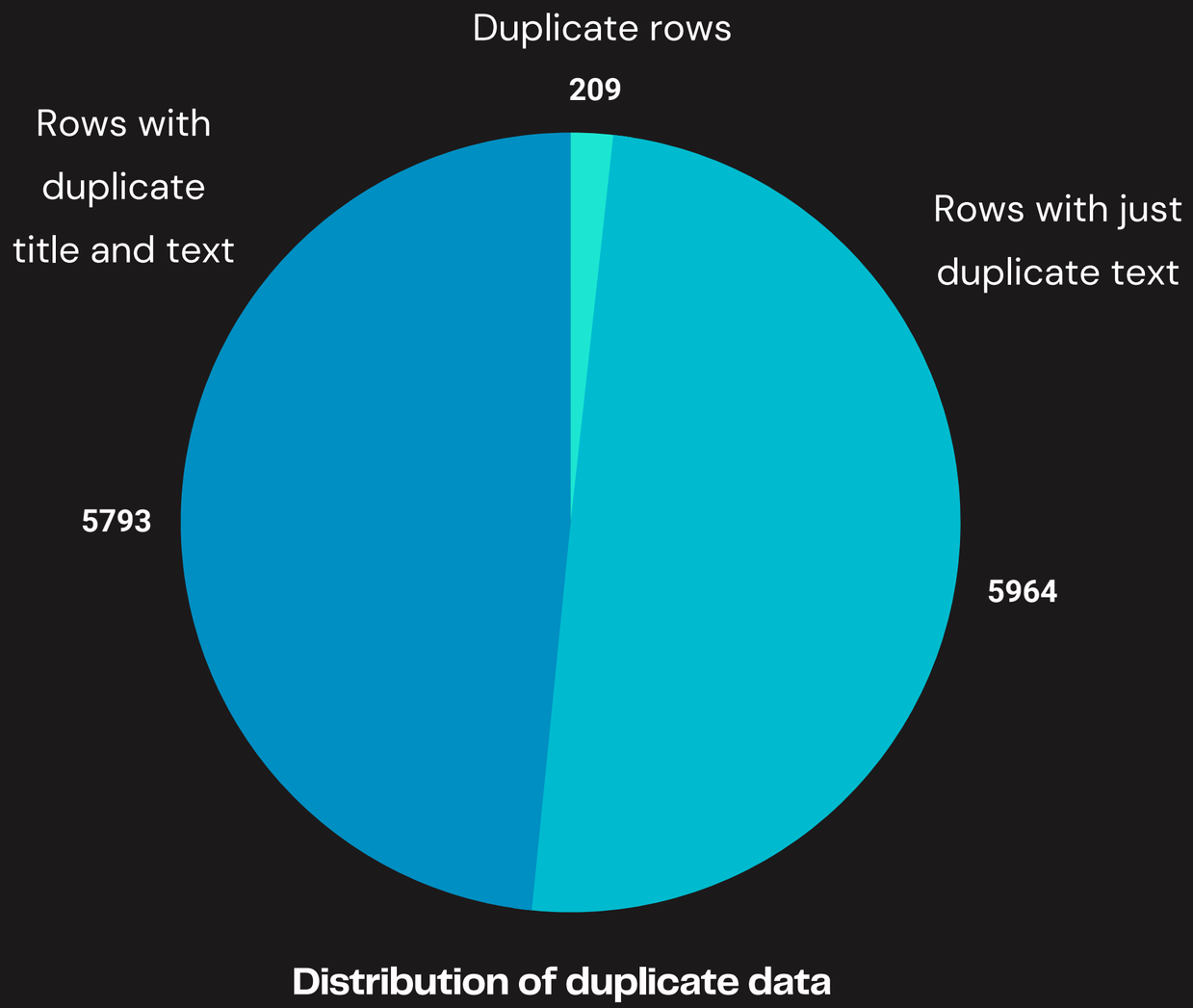
Fix: Given that the data in question look similar to what we already have in that range & because the number of instances labeled fake has already been reduced because of duplicate data, these were not removed.



Placeholder values & duplicates

It was found that there were titles being used more than once by Reuters, while changing the underlying story. Some rows maybe complete duplicates while others maybe duplicates except for the date.

Fix: There were close to 11,800 rows that have titles used multiple times. The use of a duplicate title seems to be a common practice and some articles seem to be revisions. A duplicate title is considered fine and is kept, but the duplicate rows and rows with duplicate text for the story were removed. Rows without text were also removed.



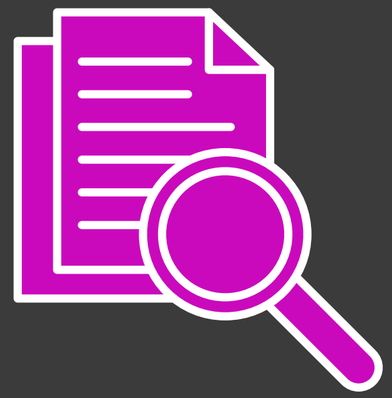
Distribution of duplicate data

Cleaning & filtering the supplemental dataset

Unlike the primary dataset, the supplemental dataset requires very minimal cleaning & filtering. Only two operation were performed on the dataset to prepare it:

1. Separate data: We only need the political news from this dataset, containing approximately 12,650 articles so that they are similar to the stories in the existing dataset. Hence, we filter out the rest.

2. Selecting appropriate fields: Although the dataset has a lot of fields, we only require the title, body text and date. Hence, we filter out the rest.

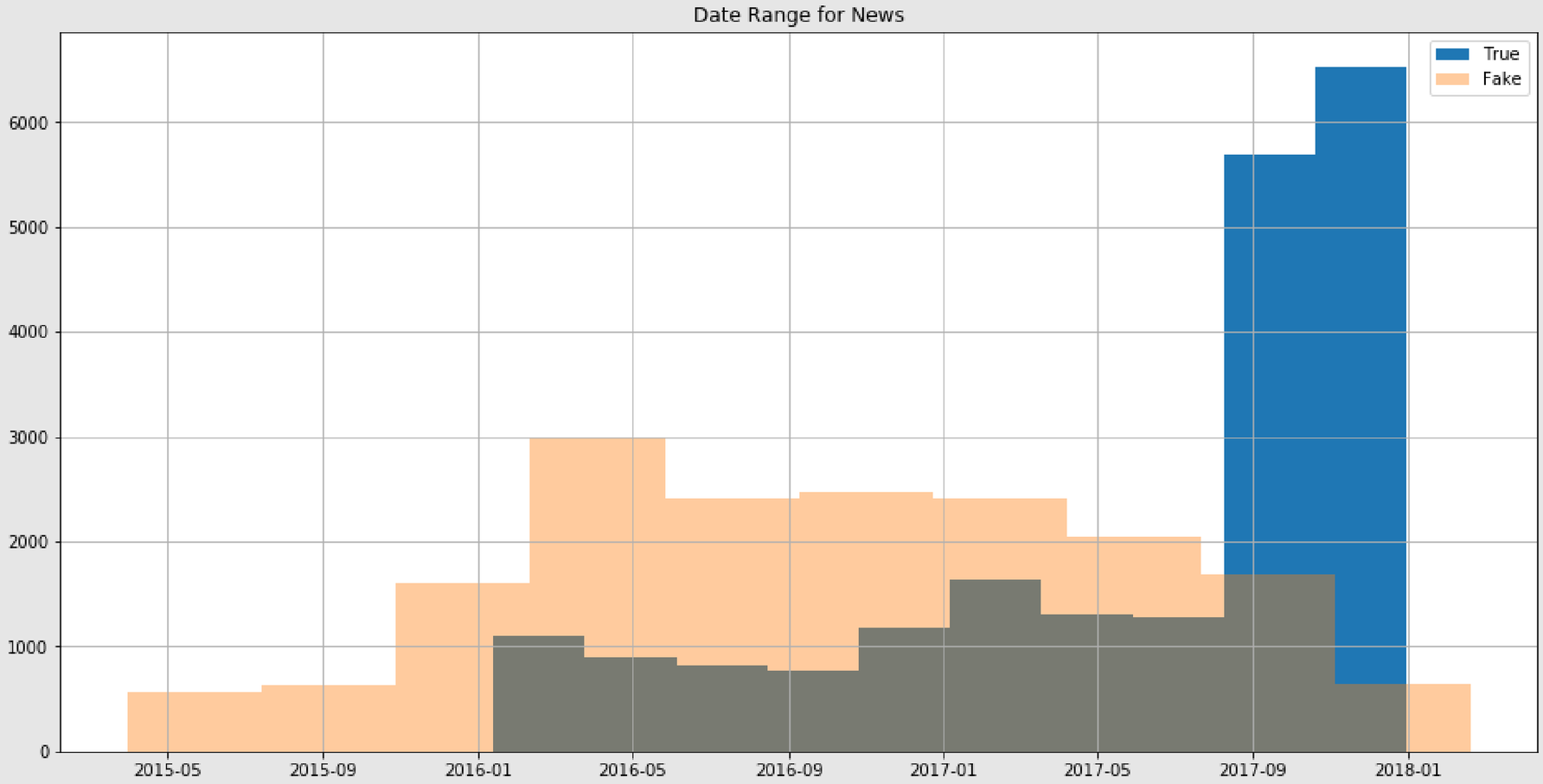


Exploratory Data Analysis

Distribution of Labels

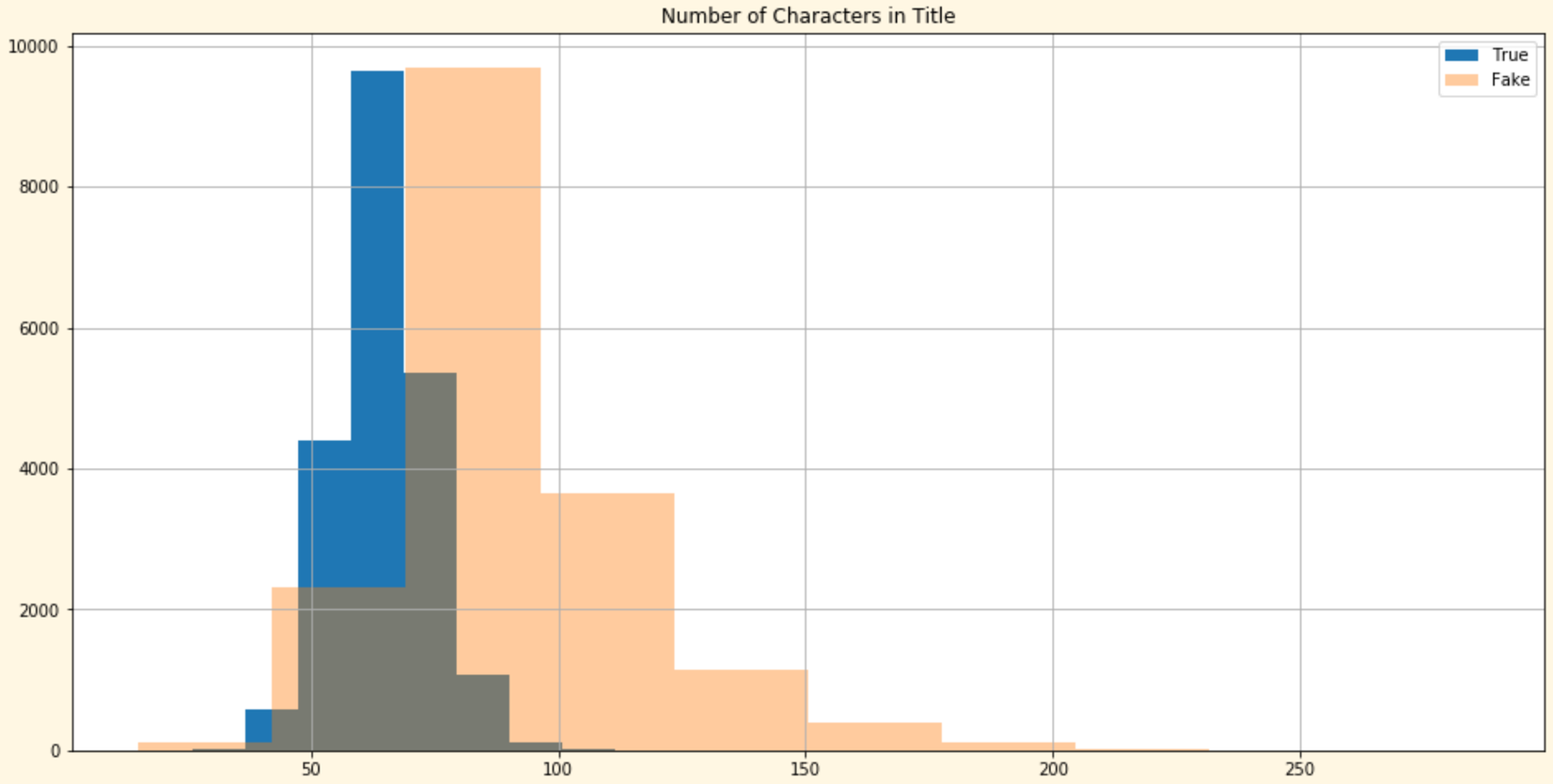
- After cleaning the data, distribution has become slightly imbalanced, but it is not significant to have a major impact on models.

Distribution of news over time



- It can be seen there is a more even distribution of news across the time frame for Fake vs. True. Given that the "news" in the dataset does not evenly cover the same events, it may give the classification models some trouble, or overfit to the dataset.

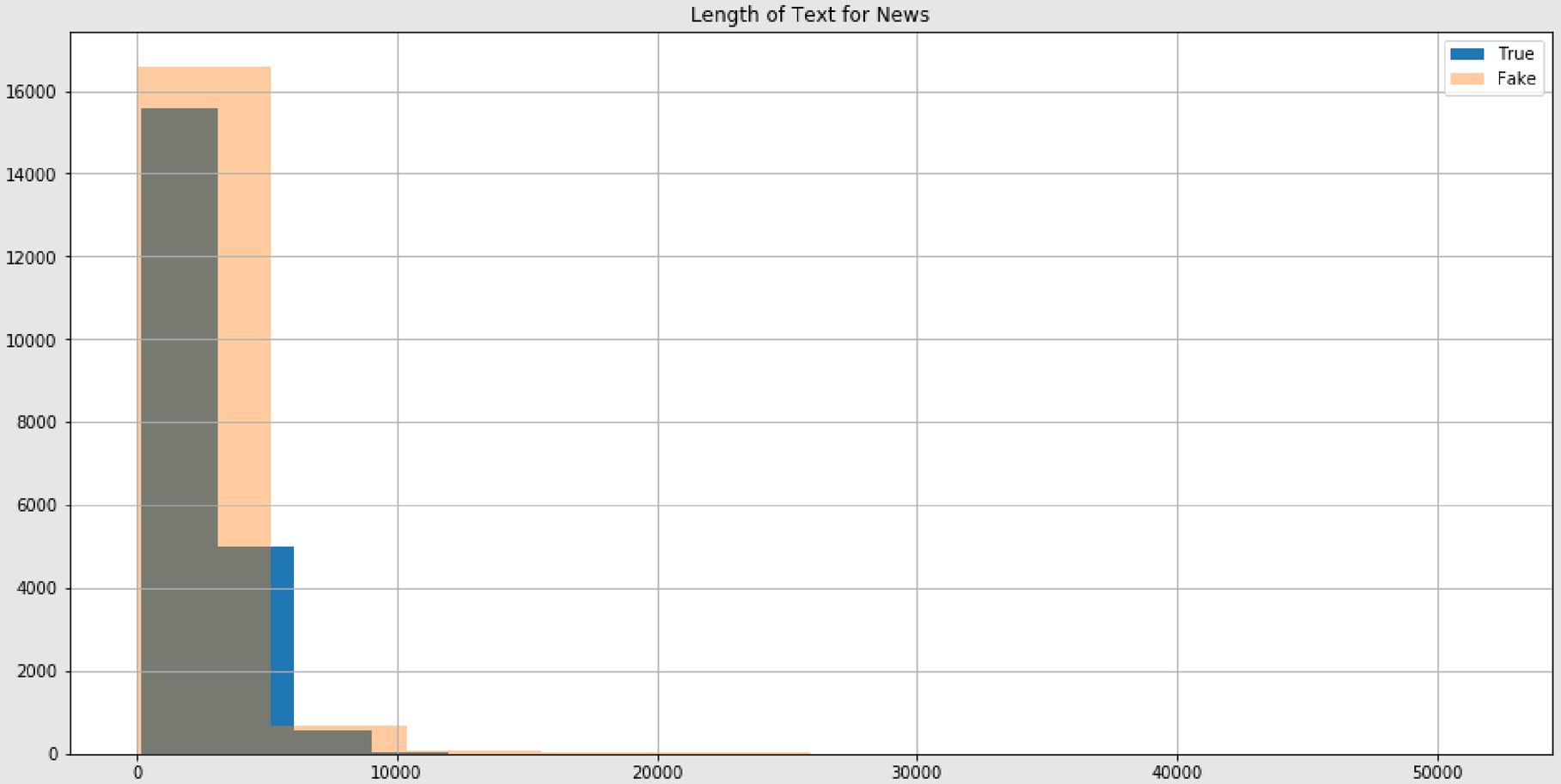
Length of title



- It can be seen fake news stories have a wider range in the length of title than True, and those Fake news titles have a higher median & mean than True ones.

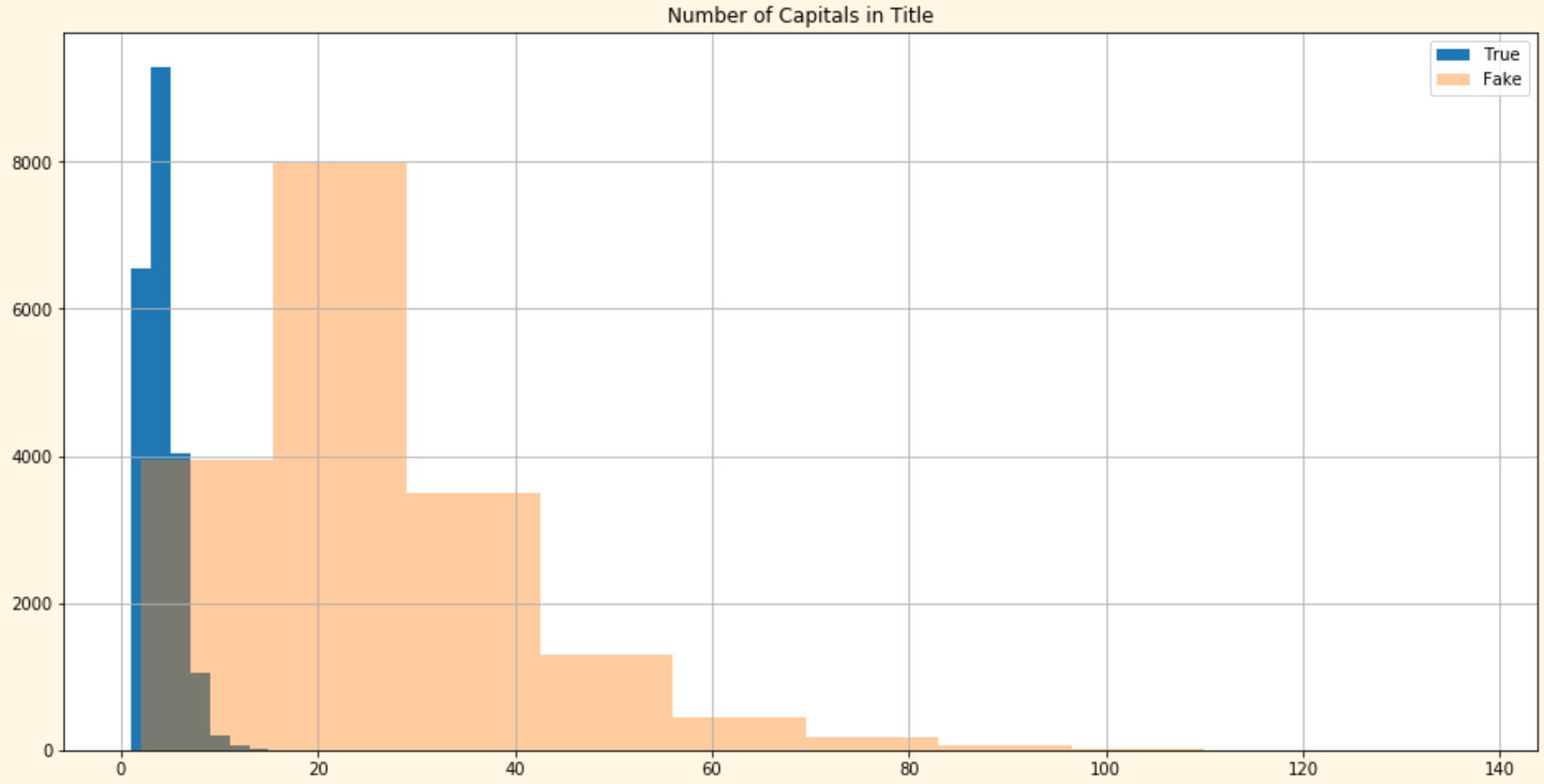
- The first quartile for Fake news is longer in length than the 3rd quartile for true news. Theoretically, labeling the news according to length of title would give better than random classification results.

Length of text



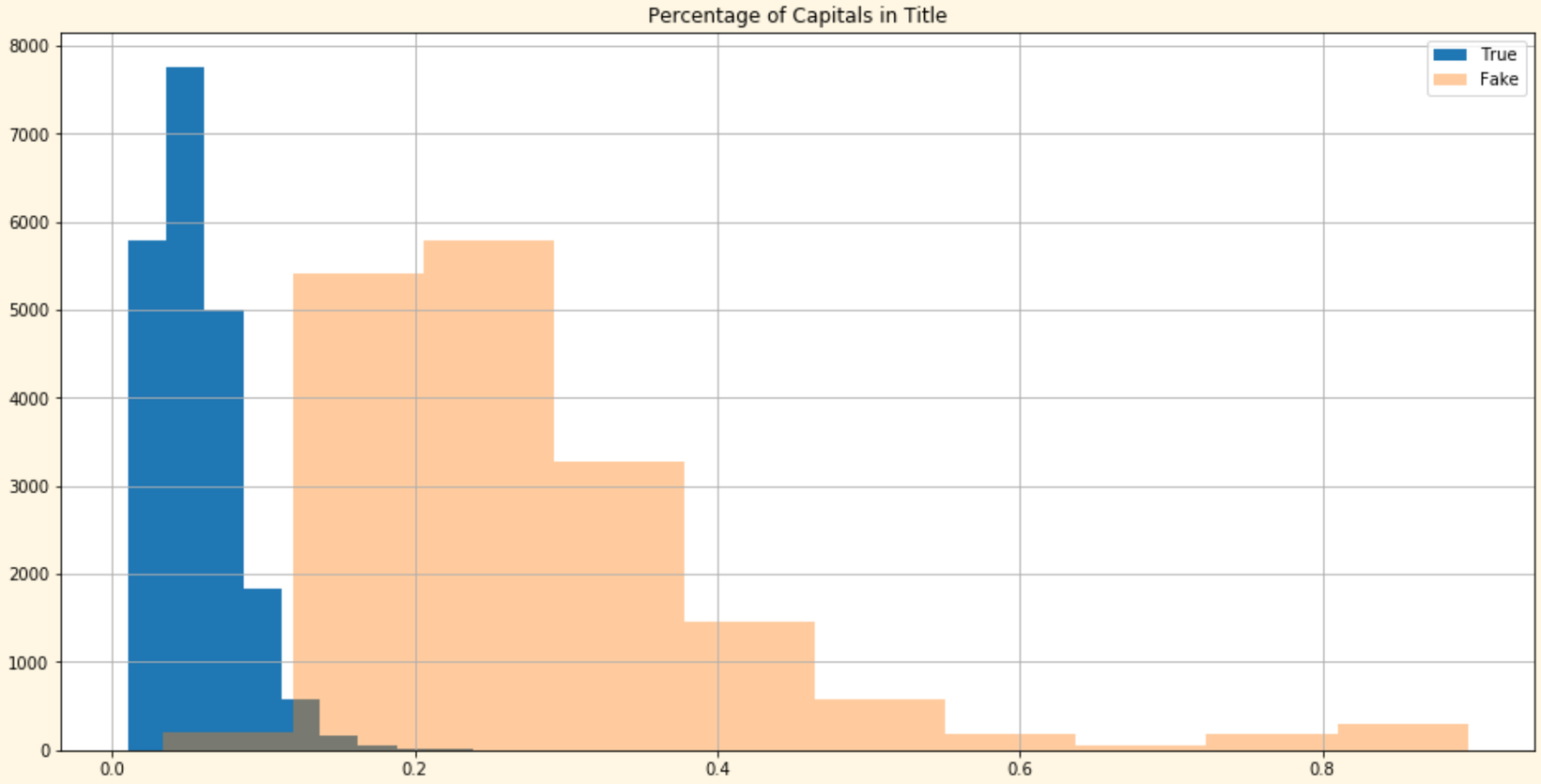
- It can be observed that there are some really long news stories (over 10,000 characters). On further inspecting the actual content of the news, it can be concluded that these are actual news stories and not invalid data.

Capital letters in title



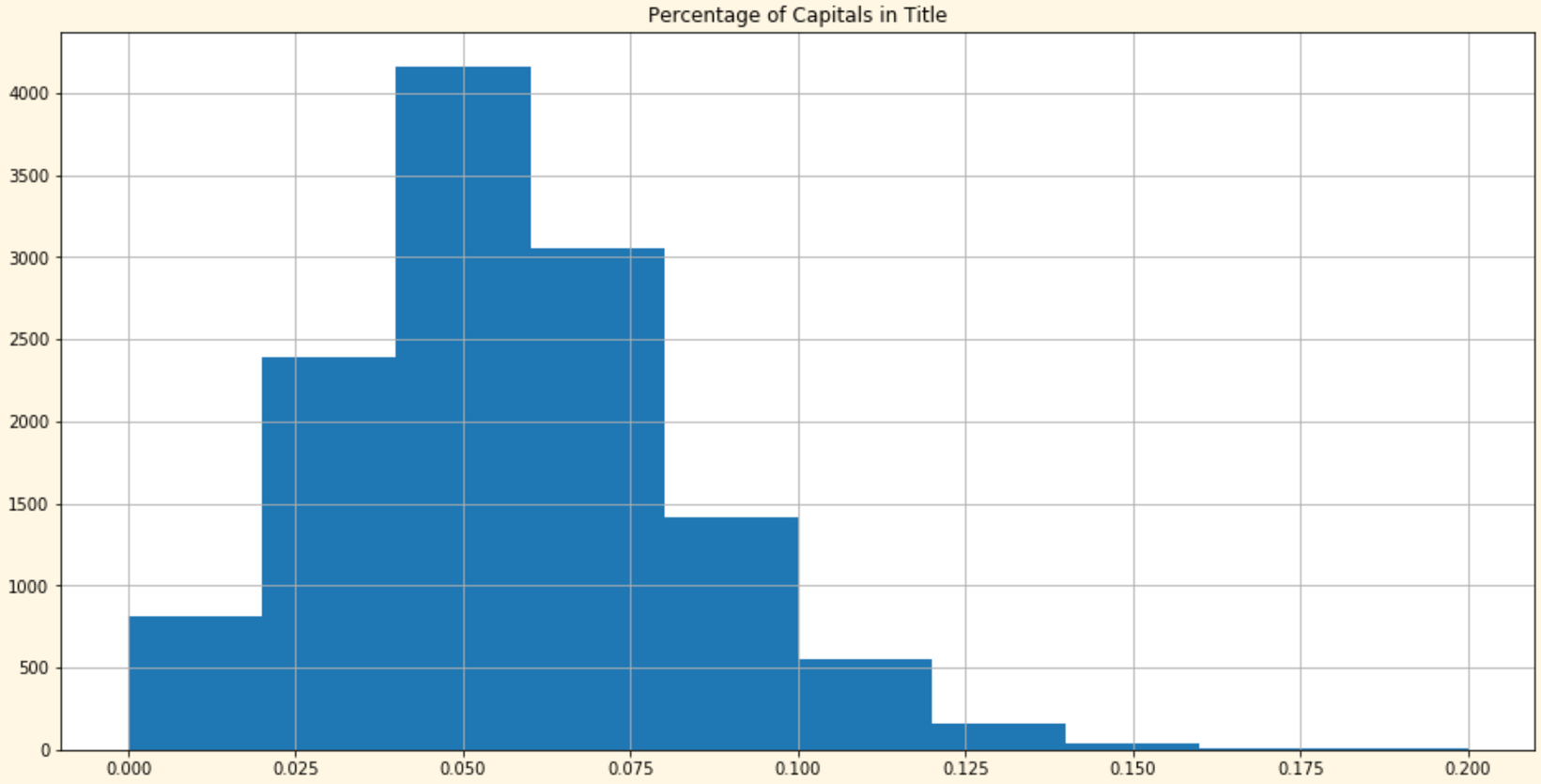
- It can be observed there are considerably more capitals in fake news compared to true. We also know that fake news instances tended to have longer titles. These two can be considered as ways to classify both for later on.

Percent of capital letters in title



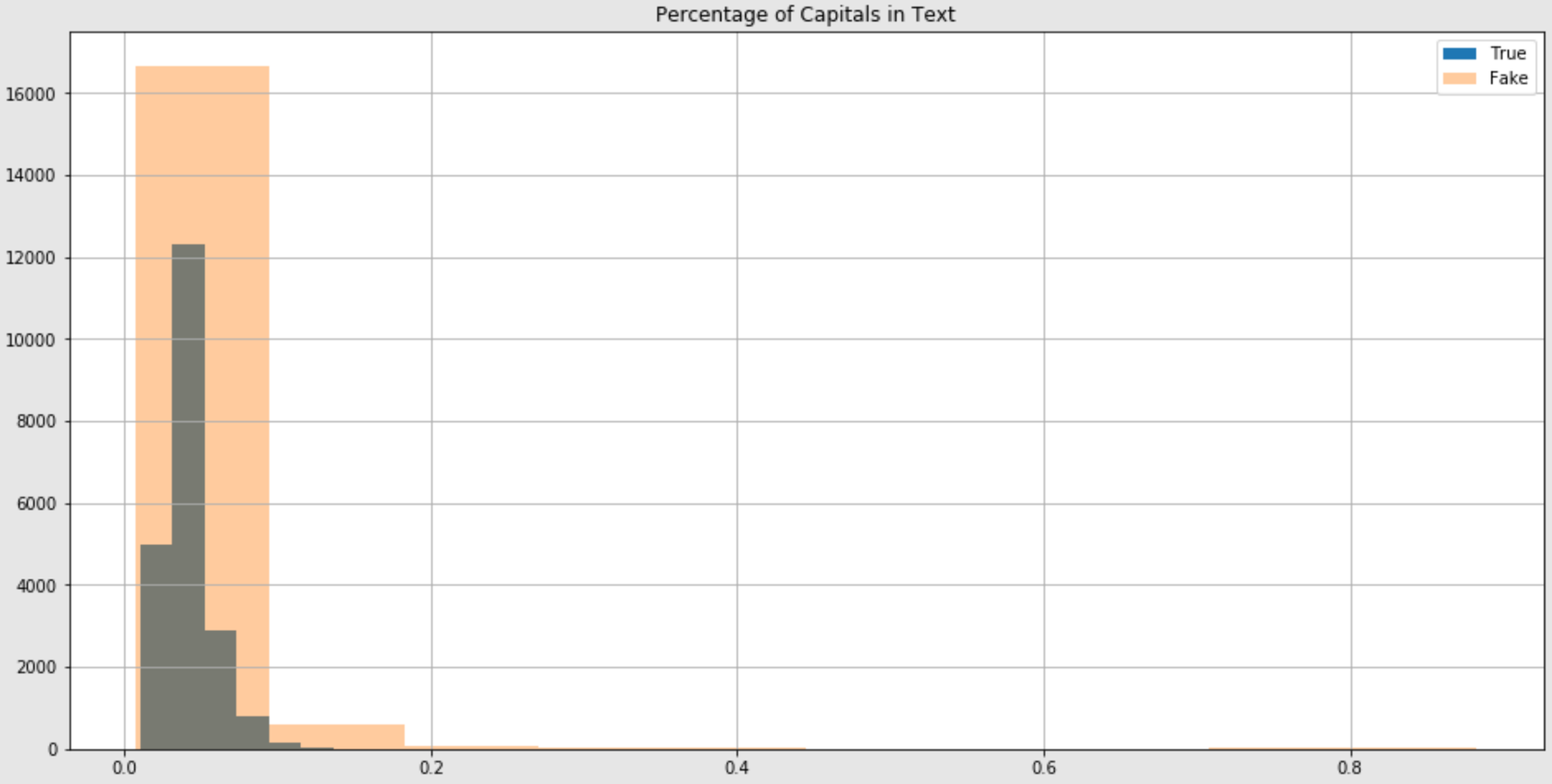
- There is less overlap when considering the percentage of capitals in the news title for Fake news compared to True. This metric would give considerably better results for classification than random guessing, and is a good candidate for a baseline model. However, it could easily be thwarted if it was used as a filter because the Fake News creators could adopt more standard capitalization standards.

Percentage of capital letters in title - supplemental dataset



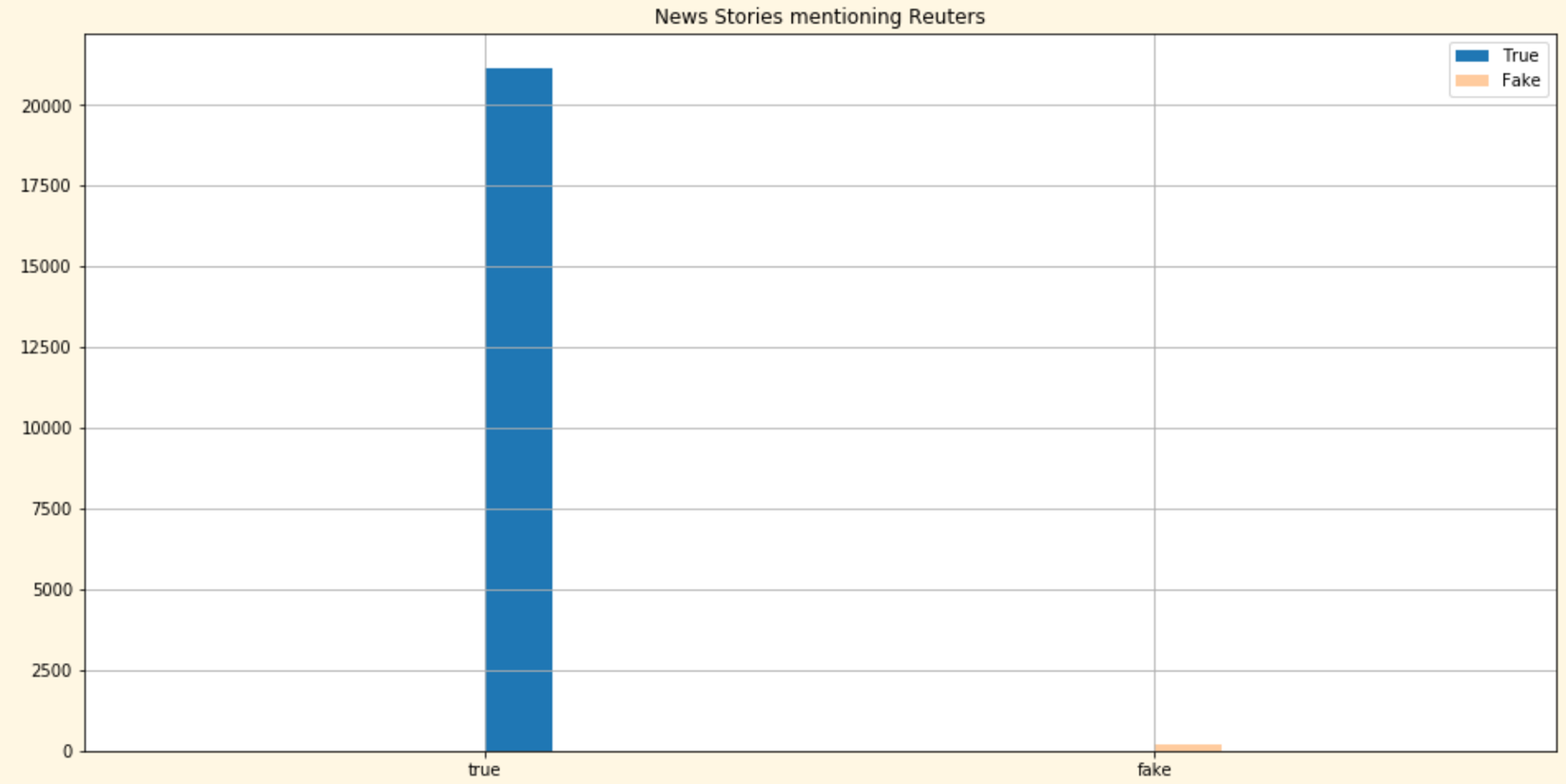
- The percentage of capitals in the titles here (supplemental dataset) do overlap with the true stories from the original dataset but it doesn't imply the heuristic would generalize.

Percentage of capitals in text



- As opposed to the title, it can be seen there is a great deal of overlap for the percentage of capitals in the text of Fake vs. True news.

News organizations mentioned in text



- It was mentioned by the dataset creators that all the True news stories are from Reuters. We only checked for Reuters with a capital R. Most but not all of the True stories include that string, plus 215 of the Fake stories reference it.

- Following this, we checked what the text for the True stories is like when they don't contain the 'Reuters' text. Interestingly there were a few stories with the True label that had the label 'IFR' even though it was mentioned all true stories are from Reuters. The stories without a news source look to follow the standard format of a Reuters news story.

Extra Observations

Apart from the info we've gathered via EDA using graphs and figures, we've also figured out a few words that seem to help classify fake and real news:

'via'

When the distribution for 'via' was checked, the following results were obtained:

fake: 7626
true: 580

It can be said "via" is highly indicative of a fake news story.

'image via'

When the distribution for 'image via' was checked, the following results were obtained:

fake: 6061

With all the posts with 'image via' being fake, it's highly indicative of that label but this may be particular to this dataset and may not generalize but it is worth noting.

'on'

Distribution:

true: 20968
fake: 15082

The use of 'on' is fairly balanced although somewhat indicative of a 'true' story.

'said'

Distribution:

true: 19855
fake: 9849

The stories containing the word 'said' are indicative of the news story being true. With twice as many of the "true" news stories containing said vs. "fake", the true ones must seem likely to be more concerned with providing quotations, or at least quotations in this style.

'you'

Distribution:

fake: 8558
true: 2401

Although only in about half of the fake news stories, it seems highly indicative of a fake news story because of the comparatively low true news stories that have the word 'you'.

Text Preprocessing

Before we proceed to modelling, we need to do some text pre-processing. We performed the following operations to our dataset to make it prepared for modelling:

1. Replace Twitter handle: While there may be some value in keeping the actual Twitter handles, we felt there would be more value in just mentioning that a Twitter handle was used instead. Hence they were replaced with '@twitter-handle' for all values.

2. Capitalization: Since words with all caps are an important way that emphasis is made online, we have kept words that are in all caps while making all the letters in other words lower case. Words of length one were made lower case regardless since there is not much emphasis lost.

3. Numbers: Numbers do not seem likely to indicate fake news, although certain dates or numbers may. One such that may have significant meaning is 9/11, which was replaced to nine-eleven so that others can be more easily removed.

4. Remove 'Reuters' from news stories: Keeping it in the news text will create an overfit model when applying it to data outside the current dataset hence it was removed.

5. Remove punctuation and single letter tokens from text: All of the punctuation tokens except for the exclamation point were removed because it seems like it may be an indicator of Fake news. All the single characters except for 'i' were removed too.

6. Remove 's: While the fake news frequently or always didn't removed the apostrophe from 's, it doesn't look like that was done to the true news. 's were removed so that it doesn't become a false indicator of true news.

7. Remove date words: It was noticed that some dates, especially 'Wednesday' came up as an important feature. Hence, all datewords were removed to generalize the dataset.

For the supplemental dataset, the text & title were first tokenized then normalized.

MODEL SELECTION

Based on our observations from tinkering around with the dataset, we proposed trying to use one non-ML based model along with four ML-based supervised classification models. They include:

(a) Heuristic model
(b) Naive Bayes (using Bag of Words)
(c) Random Forest (using Bag of Words)
(d) Random Forest (using TF-IDF)
(d) Random Forest (using TF-IDF and supplemental dataset)

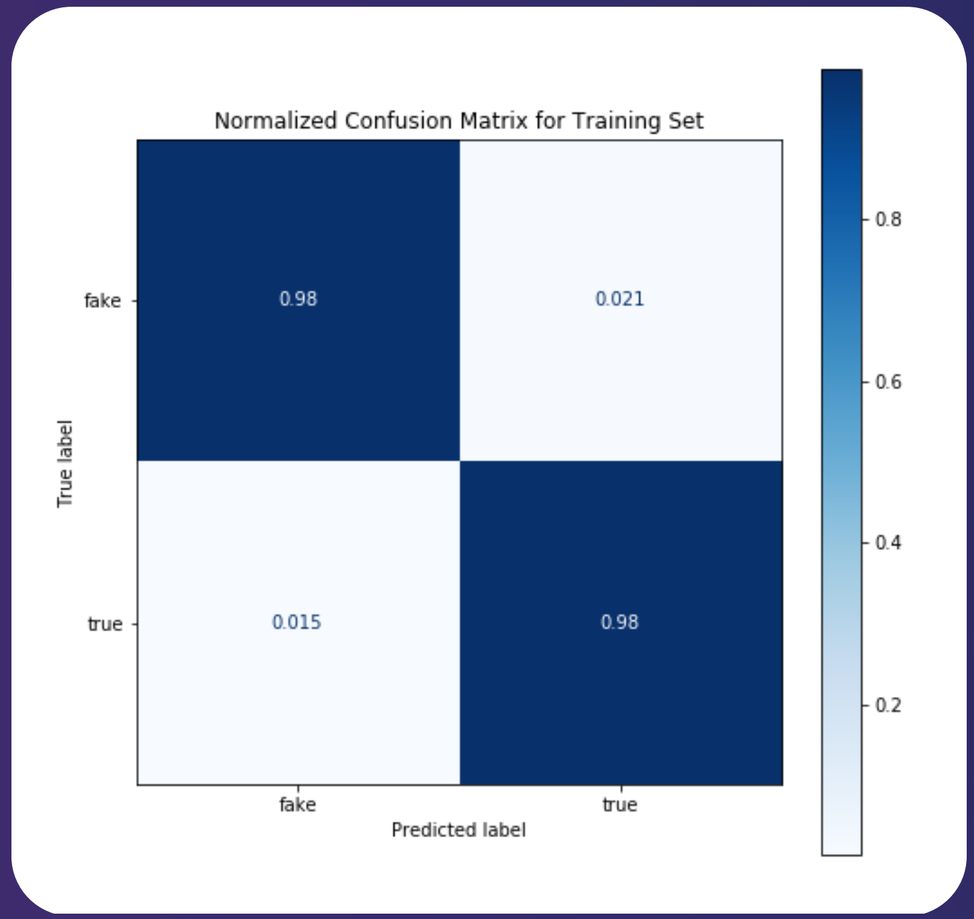
We will take a look at the results of all in the upcoming slides and then decide on a model.

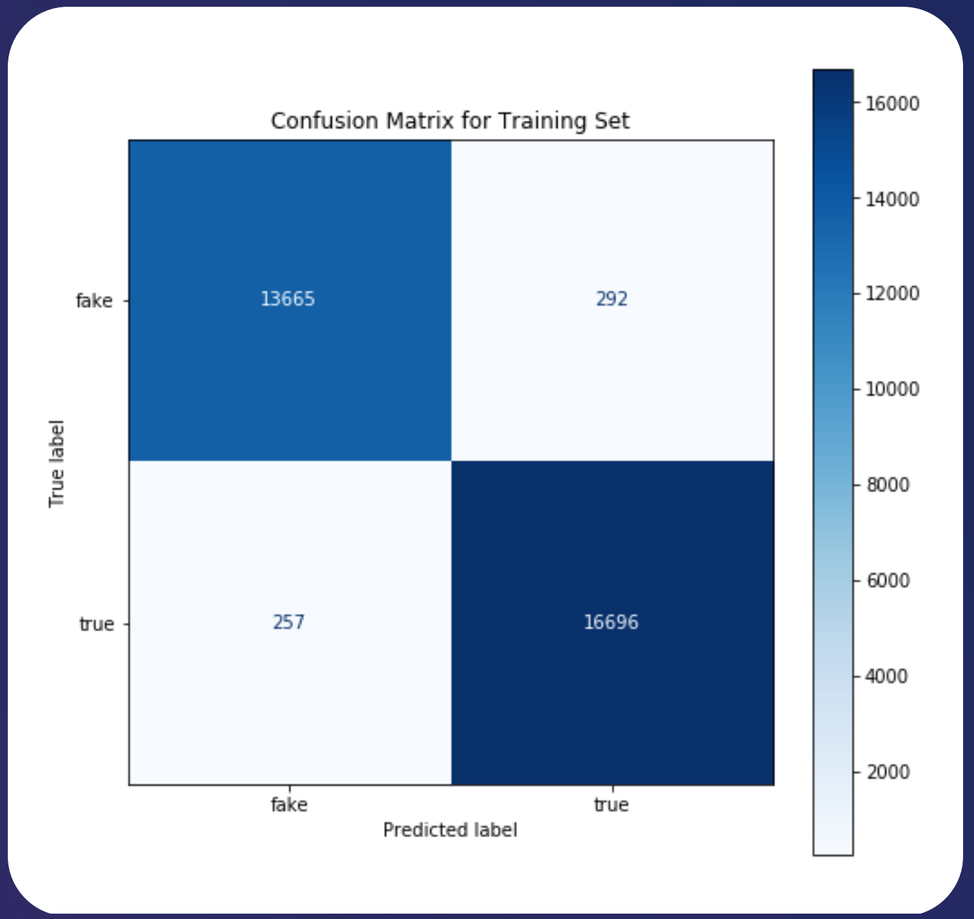
Heuristic Model

- During the EDA process, it was observed that by merely looking at the percentage of capital letters in the news story title, the fake and true news stories have very little overlap. The heuristic model is based upon this separation.

- It provides 98% accuracy and is highly balanced in its miscalculations of about 2% each although the amount of true news misclassified as fake is slightly lower than vice versa.

- However, it seems unlikely that it would generalize well and could easily be defeated if it was used as some sort of gatekeeping. This situation can be likened to the continuous updates that need do be done to a spam filter. Also, the text overfits to this dataset. There may be more fake news like the 2% that is misclassified in the wild and this test would fail.


Normalized Confusion Matrix for Training Set
fake
true
0.98
0.021
0.015
0.98
True label
fake
true
Predicted label
0.8
0.6
0.4
0.2


Confusion Matrix for Training Set
fake
true
13665
292
257
16696
True label
fake
true
Predicted label
16000
14000
12000
10000
8000
6000
4000
2000

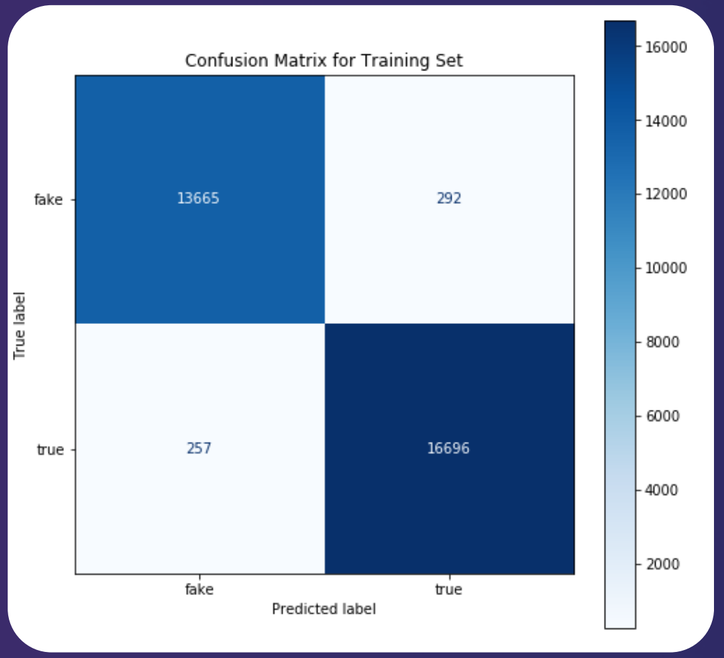
NB (using Bag of Words)

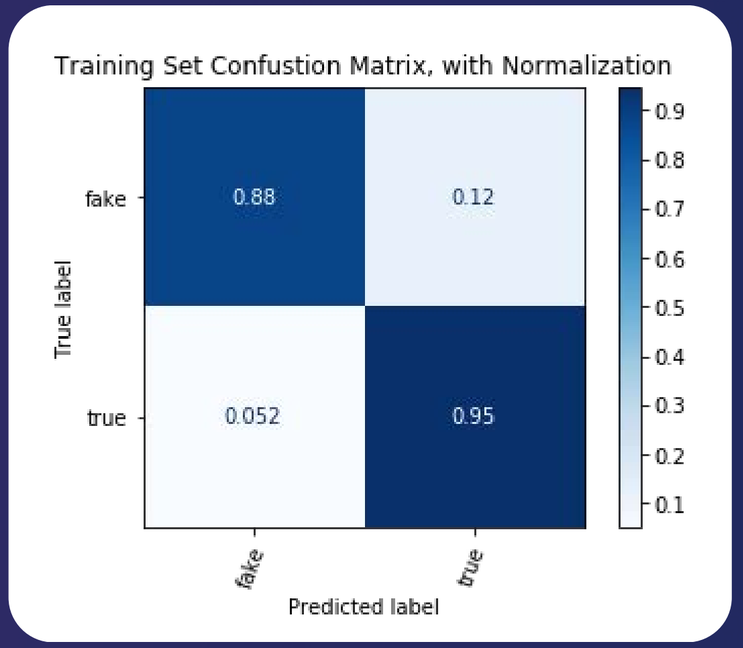
Classification Report for Training Set

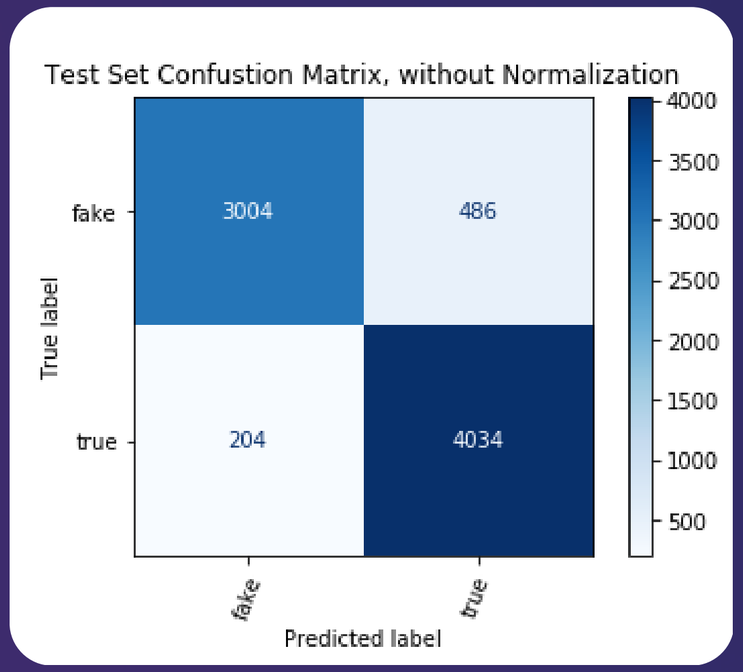
	precision	recall	f1-score	support
fake	0.96	0.95	0.95	13957
true	0.96	0.97	0.96	16953
accuracy			0.96	30910
macro avg	0.96	0.96	0.96	30910
weighted avg	0.96	0.96	0.96	30910

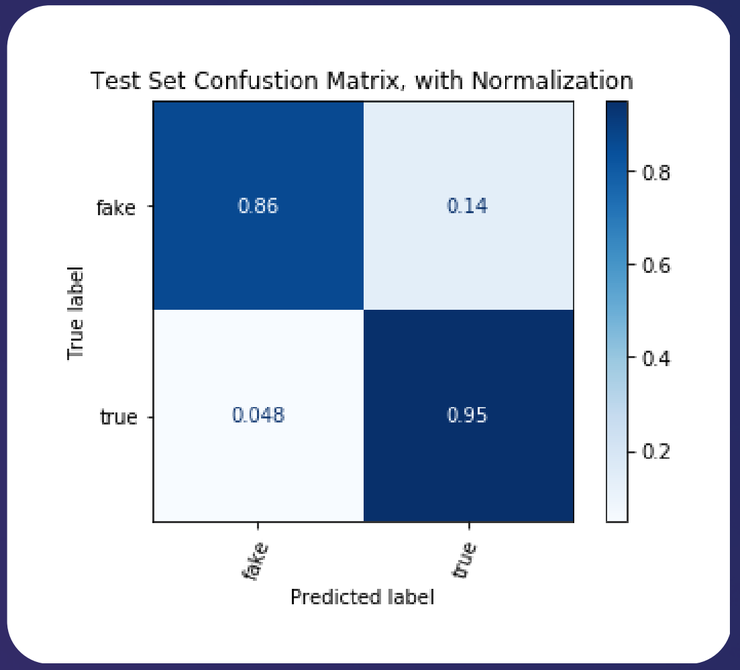
Classification Report for Test Set

	precision	recall	f1-score	support
fake	0.87	0.79	0.83	3490
true	0.84	0.90	0.87	4238
accuracy			0.85	7728
macro avg	0.86	0.85	0.85	7728
weighted avg	0.85	0.85	0.85	7728


Confusion Matrix for Training Set
fake
true
13665
292
257
16696
True label
fake
true
Predicted label
16000
14000
12000
10000
8000
6000
4000
2000


Training Set Confustion Matrix, with Normalization
fake
true
0.88
0.12
0.052
0.95
True label
fake
true
Predicted label
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1


Test Set Confustion Matrix, without Normalization
fake
true
3004
486
204
4034
True label
fake
true
Predicted label
4000
3500
3000
2500
2000
1500
1000
500


Test Set Confustion Matrix, with Normalization
fake
true
0.86
0.14
0.048
0.95
True label
fake
true
Predicted label
0.8
0.6
0.4
0.2

- This model uses just the stopwords and provide good generalization and would be useful when using the model on news stories from outside the dataset.

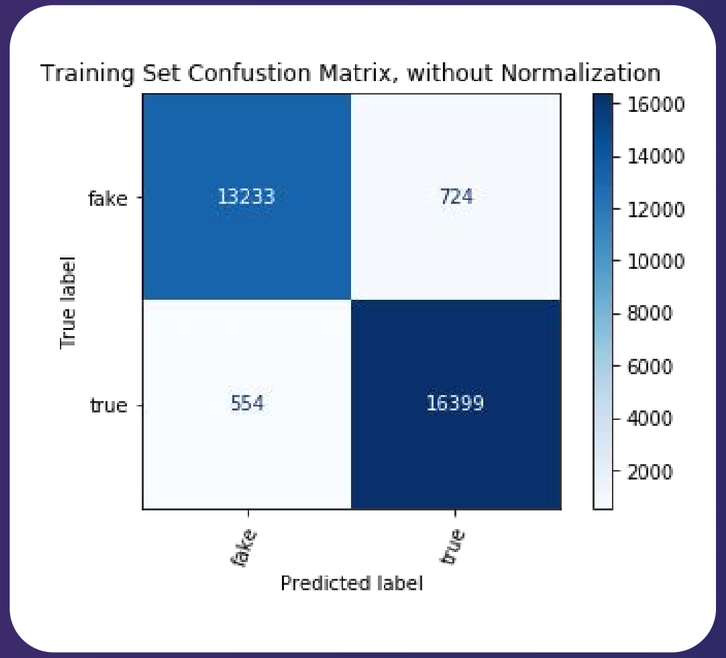
RF (using Bag of Words)

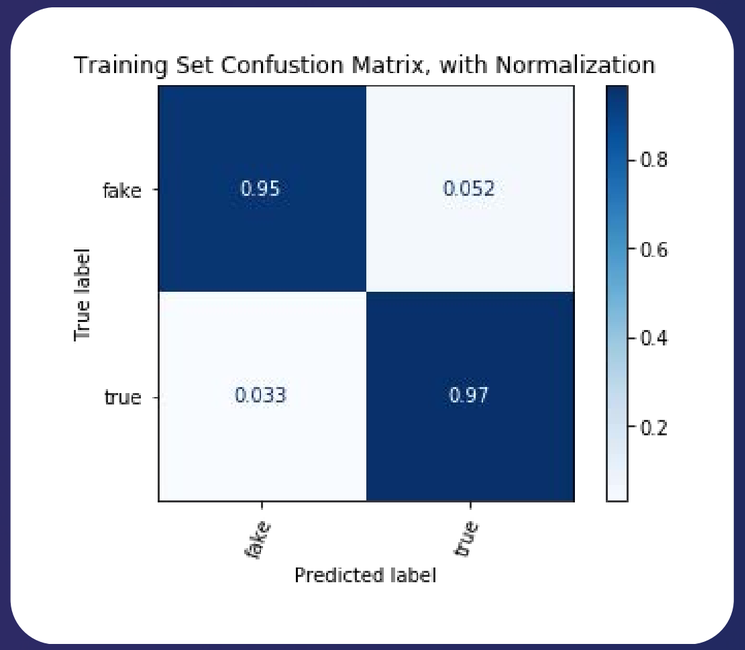
Classification Report for Training Set

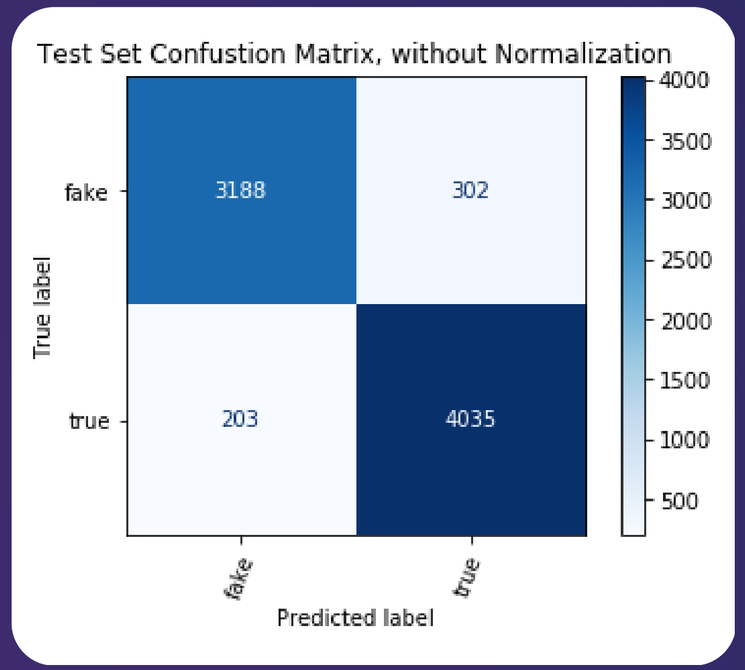
	precision	recall	f1-score	support
fake	0.96	0.95	0.95	13957
true	0.96	0.97	0.96	16953
accuracy			0.96	30910
macro avg	0.96	0.96	0.96	30910
weighted avg	0.96	0.96	0.96	30910

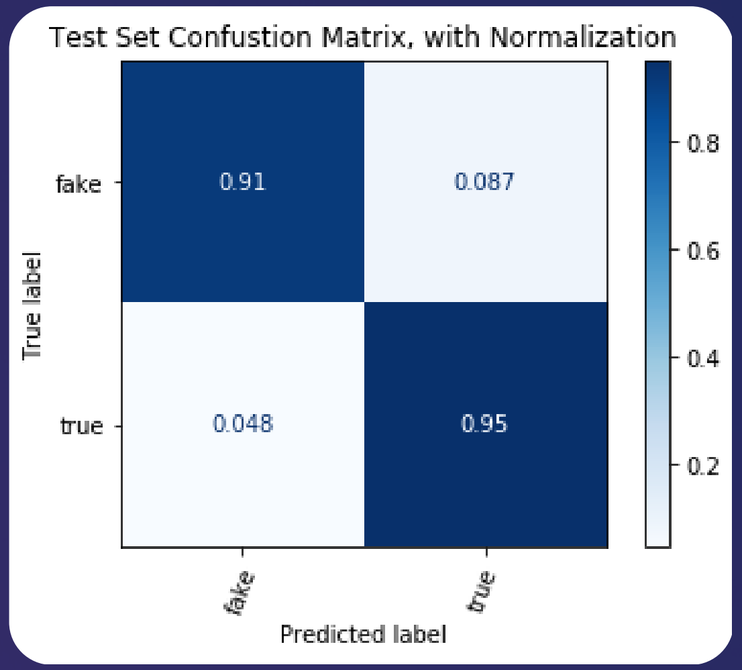
Classification Report for Test Set

	precision	recall	f1-score	support
fake	0.94	0.91	0.93	3490
true	0.93	0.95	0.94	4238
accuracy			0.93	7728
macro avg	0.94	0.93	0.93	7728
weighted avg	0.93	0.93	0.93	7728


Training Set Confustion Matrix, without Normalization
fake
true
True label
13233
724
554
16399
16000
14000
12000
10000
8000
6000
4000
2000
fake
true
Predicted label


Training Set Confustion Matrix, with Normalization
fake
true
True label
0.95
0.052
0.033
0.97
0.8
0.6
0.4
0.2
fake
true
Predicted label


Test Set Confustion Matrix, without Normalization
fake
true
True label
3188
302
203
4035
4000
3500
3000
2500
2000
1500
1000
500
fake
true
Predicted label


Test Set Confustion Matrix, with Normalization
fake
true
True label
0.91
0.087
0.048
0.95
0.8
0.6
0.4
0.2
fake
true
Predicted label

- This model uses just the list of expanded stopwords on title and text. It doesn't have the highest accuracy and F1 scores but they are still at 0.93, and using a limited set of words should create a more generalizable model.

- This model also does slightly better the the Naive Bayes model on accuracy and F1 scores when using this same stopwords only dataset.

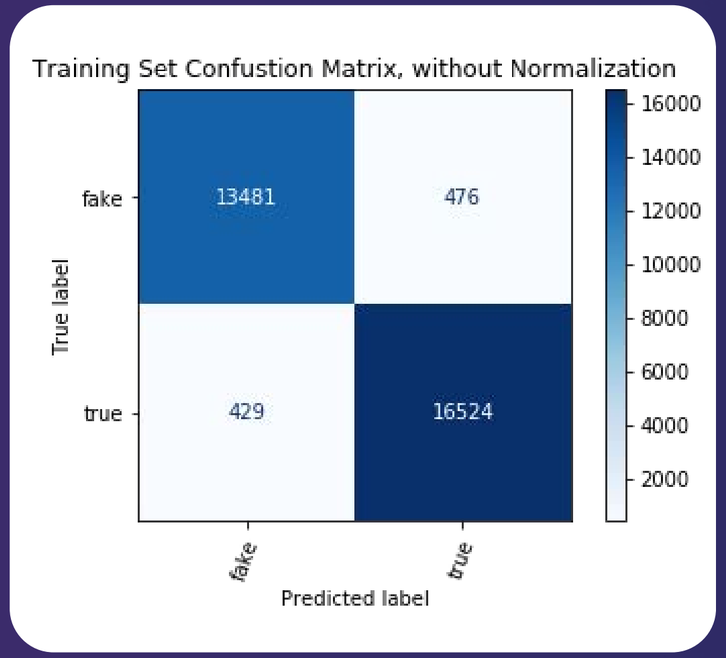
RF (using TF-IDF)

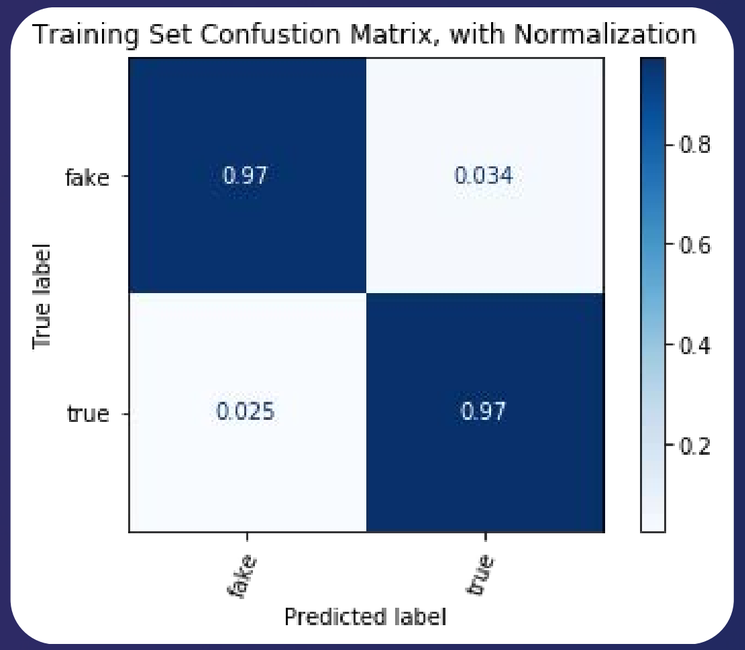
Classification Report for Training Set

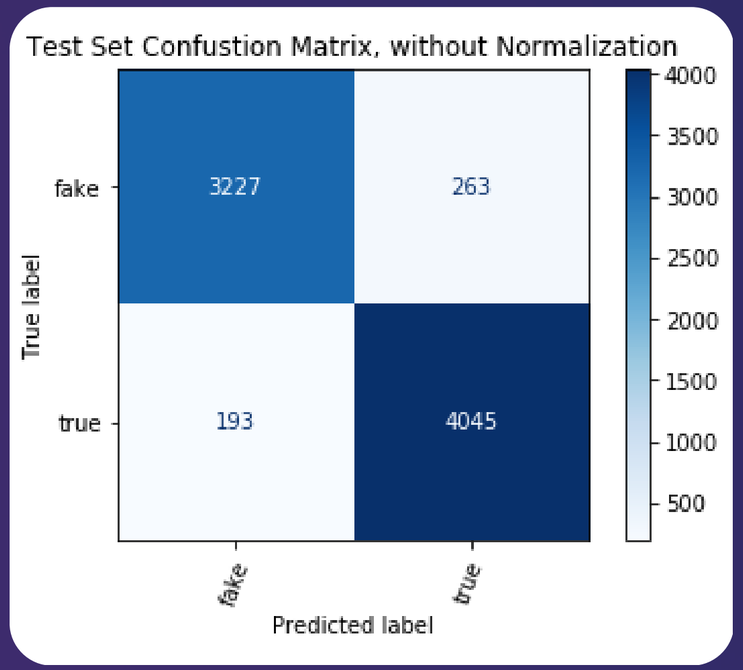
	precision	recall	f1-score	support
fake	0.97	0.97	0.97	13957
true	0.97	0.97	0.97	16953
accuracy			0.97	30910
macro avg	0.97	0.97	0.97	30910
weighted avg	0.97	0.97	0.97	30910

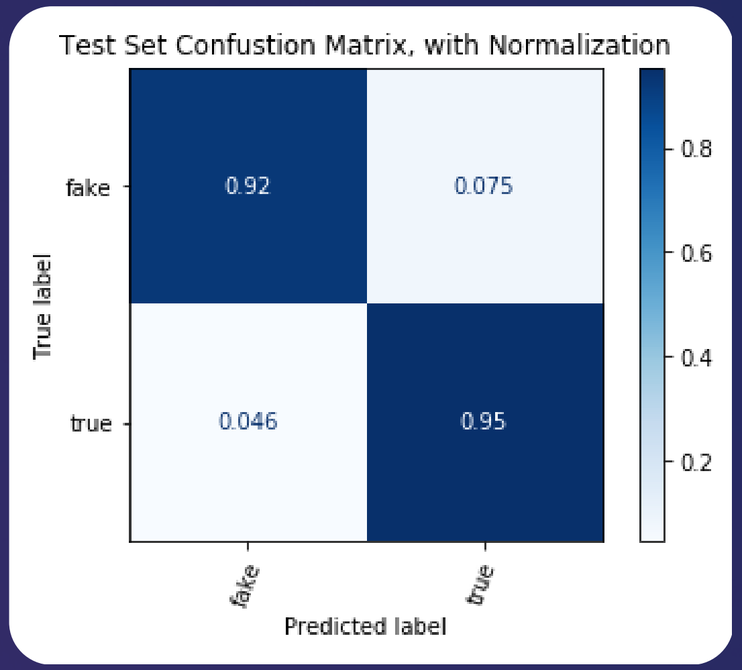
Classification Report for Test Set

	precision	recall	f1-score	support
fake	0.94	0.92	0.93	3490
true	0.94	0.95	0.95	4238
accuracy			0.94	7728
macro avg	0.94	0.94	0.94	7728
weighted avg	0.94	0.94	0.94	7728


Training Set Confustion Matrix, without Normalization
fake
true
True label
13481
476
429
16524
16000
14000
12000
10000
8000
6000
4000
2000
fake
true
Predicted label


Training Set Confustion Matrix, with Normalization
fake
true
True label
0.97
0.034
0.025
0.97
0.8
0.6
0.4
0.2
fake
true
Predicted label


Test Set Confustion Matrix, without Normalization
fake
true
True label
3227
263
193
4045
4000
3500
3000
2500
2000
1500
1000
500
fake
true
Predicted label


Test Set Confustion Matrix, with Normalization
fake
true
True label
0.92
0.075
0.046
0.95
0.8
0.6
0.4
0.2
fake
true
Predicted label

- RF-TFIDF Using tf/idf as features rather than word vectors (BOW) increased the accuracy and F1 scores by a little bit. This model is the best so far, both as far as its metrics and its ability to generalize.

- Using tf/idf rather than BOW should have helped differentiate stories by reducing the importance of some of the commonest stopwords. Surprisingly the top 5 features didn't change, but after that the order changed and some different words made it into the top 30.

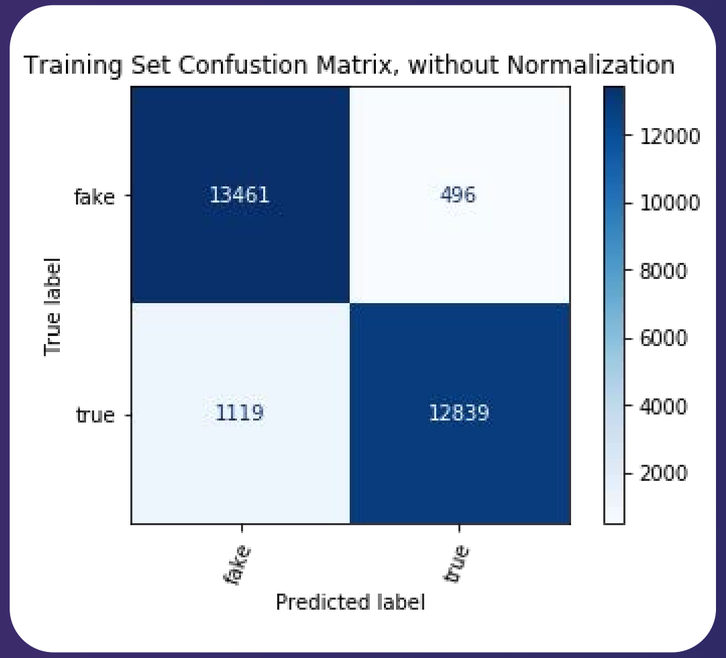
RF (using TF-IDF and supplemental dataset)

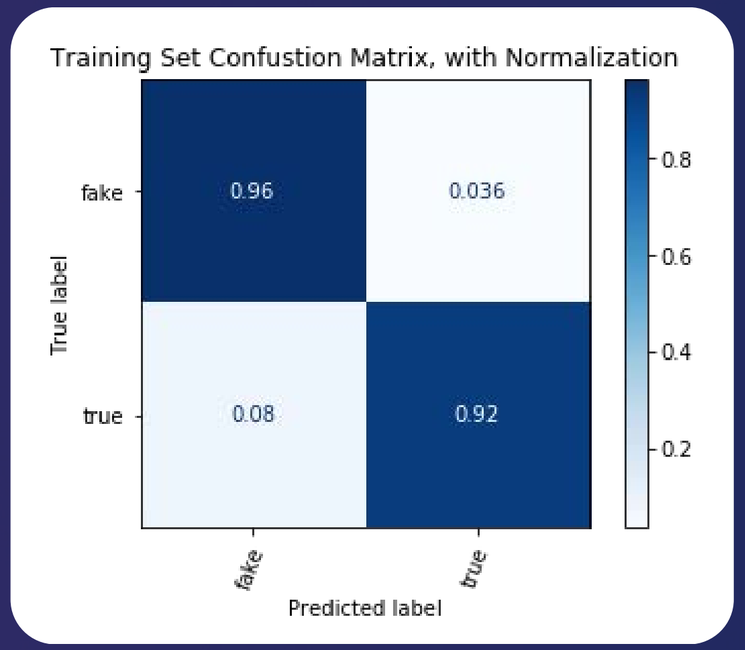
Classification Report for Training Set

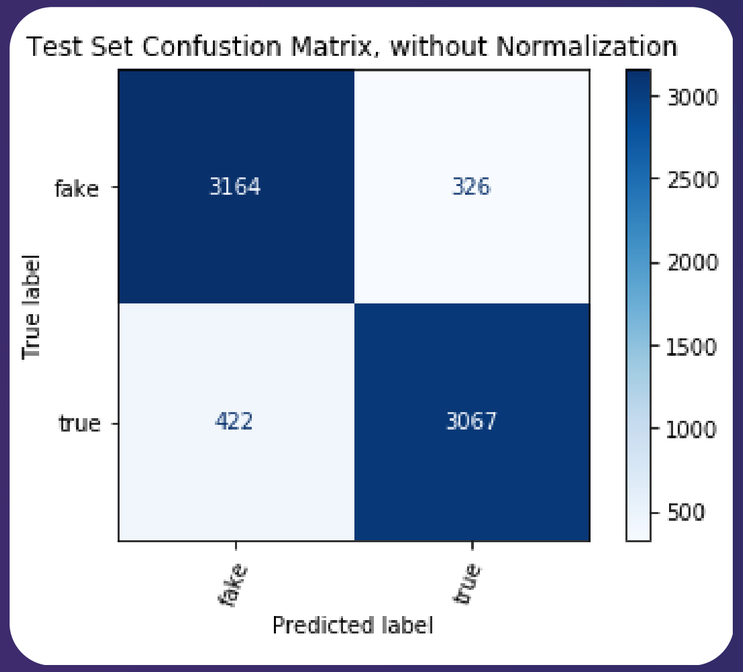
	precision	recall	f1-score	support
fake	0.92	0.96	0.94	13957
true	0.96	0.92	0.94	13958
accuracy			0.94	27915
macro avg	0.94	0.94	0.94	27915
weighted avg	0.94	0.94	0.94	27915

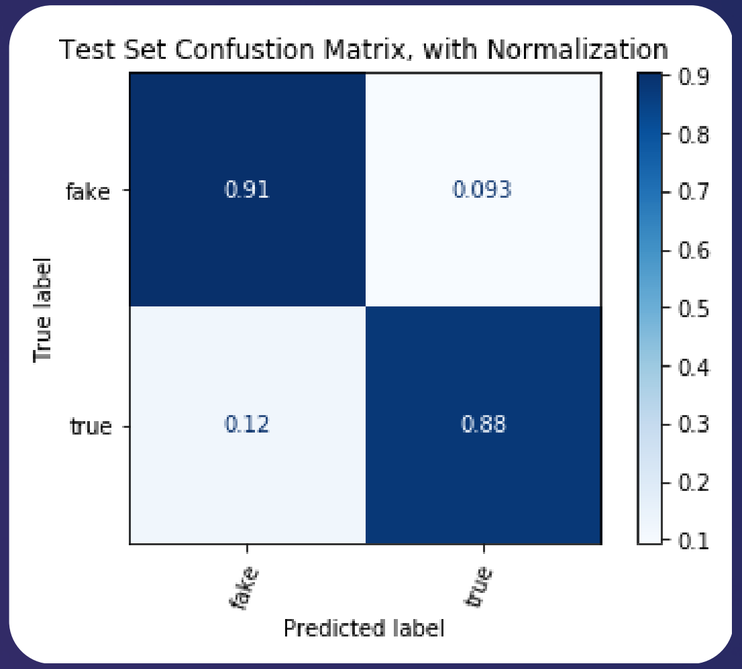
Classification Report for Test Set

	precision	recall	f1-score	support
fake	0.88	0.91	0.89	3490
true	0.90	0.88	0.89	3489
accuracy			0.89	6979
macro avg	0.89	0.89	0.89	6979
weighted avg	0.89	0.89	0.89	6979


Training Set Confustion Matrix, without Normalization
fake
true
True label
13461
496
1119
12839
12000
10000
8000
6000
4000
2000
fake
true
Predicted label


Training Set Confustion Matrix, with Normalization
fake
true
True label
0.96
0.036
0.08
0.92
0.8
0.6
0.4
0.2
fake
true
Predicted label


Test Set Confustion Matrix, without Normalization
fake
true
True label
3164
326
422
3067
3000
2500
2000
1500
1000
500
fake
true
Predicted label


Test Set Confustion Matrix, with Normalization
fake
true
True label
0.91
0.093
0.12
0.88
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
fake
true
Predicted label

- The accuracy and F1 scores for this model are somewhat worse than the model trained on the original dataset, 0.89 here vs. 0.94, but this model should generalize better because the news stories labeled true came from more than one news source.

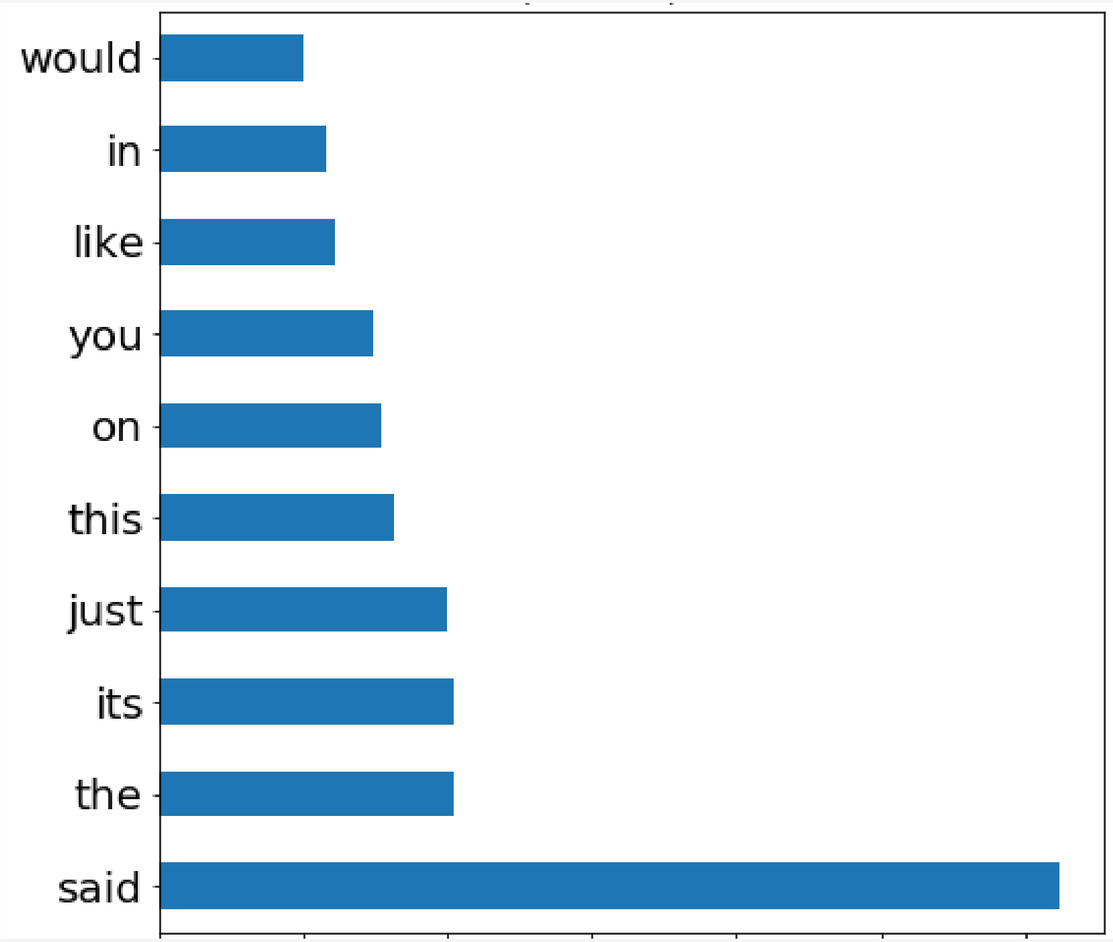
- The list of important features changed compared to what the model found for the previous version of the dataset.

Conclusion

- The model that we found to be the most generalizable while still returning an accuracy of 89% and F1 scores of 0.89 for both fake and true, was the random forest classifier that only considered words from a stopwords list and used the supplemental dataset.

- By only looking at the stopwords, the people, places, organizations, dates, jargon, and other situation specific references were removed which would make the news more generalizable because the classifications would not be dependent on those removed elements.

- While this model has a lower accuracy and F1 scores than the heuristic, it offers a more generalized solution that can be expected to perform well on news stories not in the dataset, because the news stories that have been shared online may have come from different sources or writers than the ones from this dataset.



Top 10 most important stopwords for classification according to the best performing random forest classifier.

Dataset Citation

The following citations were requested by the authors of the primary dataset:

[1] Ahmed H, Traore I, Saad S. “Detecting opinion spams and fake news using text classification”, Journal of Security and Privacy, Volume 1, Issue 1, Wiley, January/February 2018.

[2] Ahmed H, Traore I, Saad S. (2017) “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham (pp. 127- 138).

Kaggle link: https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

The following citation was requested by the author of the supplemental dataset:

[3] Sameed, H. (2019, June). Guardian News Dataset, Version 1.

Kaggle link: https://www.kaggle.com/datasets/sameedhayat/guardian-news-dataset

Kudos to these authors for providing their high quality work for free!


THANK
YOU