




DATASET DESCRIPTION

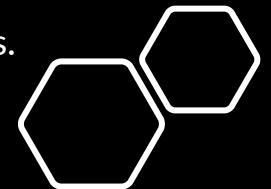
- ❖ In this project, a large **dataset** for **fake** news detection provided by using social media news and its related comments from Reddit.
 - ❖ The **dataset** consists of **69396** records (of which **74** percentage are labelled as real) by three different fact-checking sites (**Snopes**, **PolitiFact** and **Emergent**).
 - ❖ Given a record in textual format, our goal is to automatically detect whether it is fake or not.
 - ❖ The **dataset** is in json format.
 - ❖ The **dataset** is available on Kaggle (<https://www.kaggle.com/datasets/deepnews/fakenews-reddit-comments>).
- 

	label	reddit_comments	researched_by	text	title	url
0	0	[]	snopes	analyze videos growth watch videos growth sinc...	vidinfo	http://www.vidinfo.org/video/67155269/jeremy-m...
1	1	[]	snopes	last week current administration missed point ...	editorial misinterpreted toon unpatriotic	http://pittnews.com/30440/archives/editorial-m...
2	0	[]	snopes	email protected member male join date jul 2001...	general health message board	http://www.healthboards.com/boards/general-hea...
3	1	[]	snopes	urban outfitters sunk new low vintage kent sta...	urban outfitters sorry selling kent state swea...	http://gothamist.com/2014/09/15/urban_outfite...
4	1	[]	snopes	santa goes many names santa claus st nick kris...	story santa claus	http://www.englishteachermelanie.com/canada-fu...
...
69391	0	[{'created_utc': 1370998152, 'label': 0, 'auth...	snopes	donate stuff 101 places clutter good book joy ...	donate stuff 101 places clutter good miss mini...	http://www.missminimalist.com/2011/04/where-to...
69392	0	[]	snopes	update 3272012 painting depicts obama burning ...	finally obama original birth certificate surfaces	http://geopolitics.co/2012/03/21/finally-obama...
69393	0	[]	snopes	take selena movie quiz need go back watch movi...	selena quintanilla news 2014	http://loveselena.com/News/2014.html
69394	0	[]	snopes	robert well know love ya wants love do ray rob...	tropes q z everybody loves raymond	http://tvtropes.org/pmwiki/pmwiki.php/Everybod...
69395	0	[{'created_utc': 1436457724, 'label': 0, 'auth...	snopes	children crayons marketed colorful characters ...	child coloring asbestos	http://www.scientificamerican.com/article/is-y...

69396 rows x 6 columns

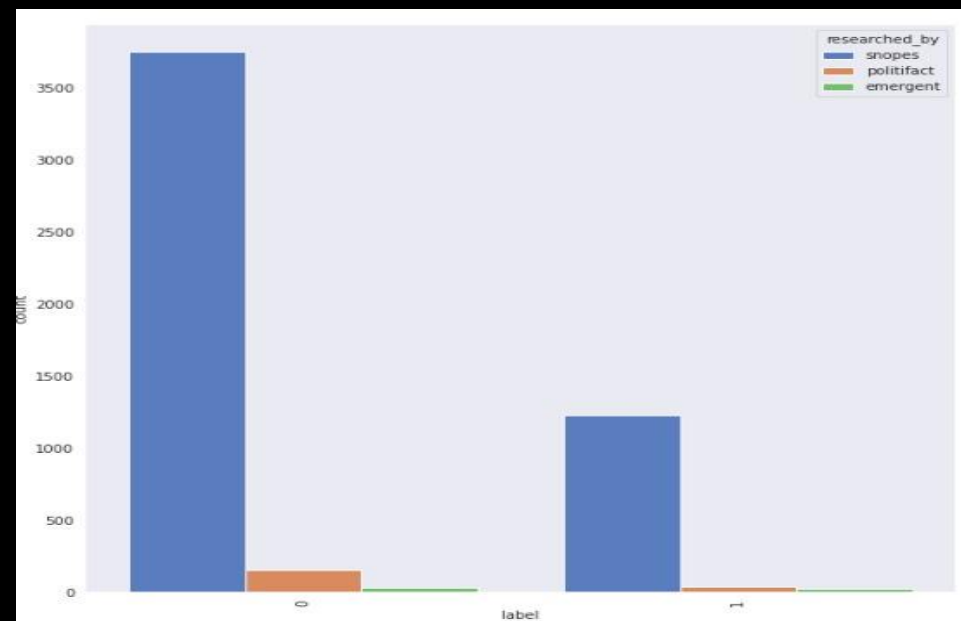
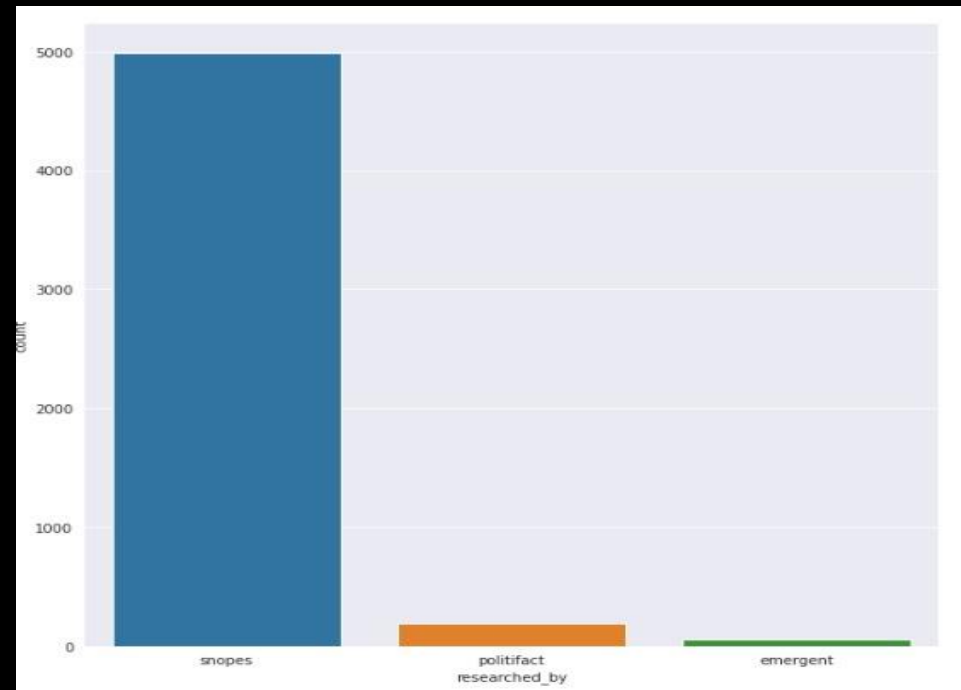
DATASET DESCRIPTION

- This dataset has six columns,
 - Title:** this represents the title of the news.
 - Researched_by:** this represents the name of the author who has written the news.
 - Ridded_comment:** contain comments about the news.
 - URL:** contain url of the news.
 - Text:** this column has the news itself.
 - Label:** this is a binary column representing if the news is fake (1) or real (0).



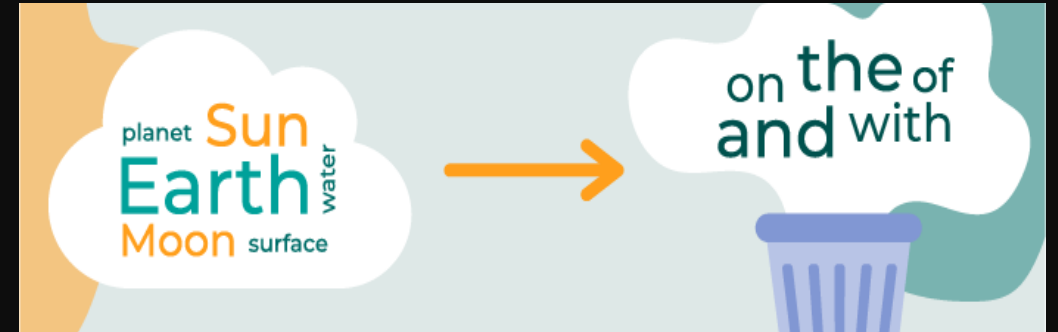
DATASET DESCRIPTION

- There are **news** checked from 3 **fact-checking** sites. We have the largest number of news from **Snopes** site. Let us dig further into this.



DATA PREPROCESSING

- For the **text** column we will make:
 - Expand Contractions
 - Lower Case
 - Remove Punctuations
 - Remove words and digits containing digits
 - Remove Stopwords
 - Stemming
 - Remove White spaces
 - Remove tags
- For the **riddet_comment**:
 - Extract new feature represents the number of comments for each news



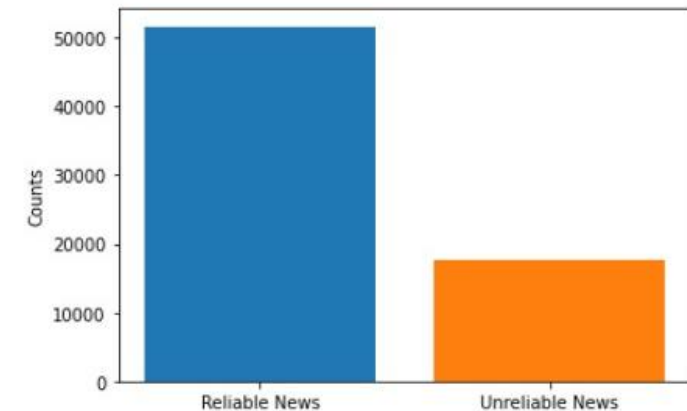
DATASET STATISTICS

After displaying data info, we observed that:

- 'label' column has int datatype
- and the other columns have object datatype.

After checking **missing** values, we observed that:

- reddit comments that equal to **64161**
- the other column have **0** missing values.



```
[17] #display label categories & their counts  
df.label.value_counts()
```

```
0    51625  
1    17771  
Name: label, dtype: int64
```

```
#check missing values  
df.isnull().sum()
```

```
label                0  
reddit_comments      64161  
researched_by        0  
text                 0  
title                0  
url                  0  
dtype: int64
```

DATA PREPROCESSING

- In data processing, we will focus on the **text** column on this data which contains the news part.
- We will modify this **text** column to extract more information to make the model more predictable.

text
analyze videos growth watch videos growth sinc...
last week current administration missed point ...
email protected member male join date jul 2001...
urban outfitters sunk new low vintage kent sta...
santa goes many names santa claus st nick kris...
...



1. First Question:

- Is the ratio of fake news in Emergent fact-checking site significantly higher than the ratio of fake news by all other fact-checking sites?

- ❖ **Motivation:**

- The benefit of answering this question is helping the public using social media to determine which sites that are used to detect **fake** news are more credible than others.

- ❖ **Approach**

- Calculating ratio of news that checked by each fact-checking site are fake.

- ❖ **Findings:**

- From calculating the ratio for the 3 fact-checking sites, we found that:

- The ratio for **snopes** is **0.25**
 - The ratio for **politifact** is **0.166**
 - And the ratio for **emergent** is **0.441**

2. Second Question

- Predicting the number of user comments for each news.

❖ Motivation:

The benefit of answering this question is helping knowing that which type of news users can interact with it and may be indicative of its importance or impact.

❖ Approach:

We extracted the length of comments of each news as a new feature, then used it as label to predict later the expected number of comments on each news and applied this through 2

scenarios:

- 1- using the whole data (69396 rows) that has missing values in reddit_comments column assuming that the remained news have no comments (0)
- 2- using part of the data that has comments on each news without the ones that has missing values assuming that their comments are missed, so delete them, and use that 5235 rows to train/test our model.

❖ Findings:

Until now we check the models to reach to the best performance

3. Last Question:

- Predicting fake news with/without URLs.

Motivation:

The benefit of this question that, knowing the type of url whether it contain fake news or not. So, the user can distinguish between them. And that if the user know that the url was fake for one news we may predict that the other news that has the same link is also fake.

Approach:

We tried to predict fake news in two different scenarios:

- 1: predicting it by some features but url.
- 2: predicting it by the same some features including the url.

Findings:

Without using url feature , we obtained better performance for predicting the fake news than using it.



THE RES. OF OUR MODELS WITH URL FEATURE:

Models	Test Accuracy
Logistic Regression	78%
Multinomial Naïve Bayes	79%
Passive Aggressive Classifier	77%



THE RES. OF OUR MODELS WITHOUT **URL** FEATURE:

Models	Test Accuracy
Logistic Regression	79%
Multinomial Naïve Bayes	79%
Passive Aggressive Classifier	79%

MODELING

- After **Vectorization**, we **split** the **data** into **test** and **train** data we will use:
 - Logistic Regression,
 - Linear Regression
 - k-nearest neighbors
 - Naive-Bayes,
 - Decision Tree,
 - and Passive-Aggressive Classifier.
 - Support Vector Classification

LIMITATIONS OF OUR WORK:

- Time taken by the **models**.
- Ram required to train the models.
- The quality of data that limits the model's **accuracy** to a specific range.

CONCLUSION

- Fake news detection techniques can be divided into those based on style and those based on content, or fact-checking. Too often it is assumed that bad style (bad spelling, bad punctuation, limited vocabulary, using terms of abuse, ungrammaticality, etc.) is a safe indicator of fake news.
- We tried to do different preprocessing on our data to prepare it for the model ...
- We tried to use different models with the data after preprocessing to show the accuracy
- We tried to achieve the best accuracy with models

