POFMA TEAM

Wednesday 15/09/2021 Newsletter Edition: 002

“Fake News Detection of COVID-19 Tweets”

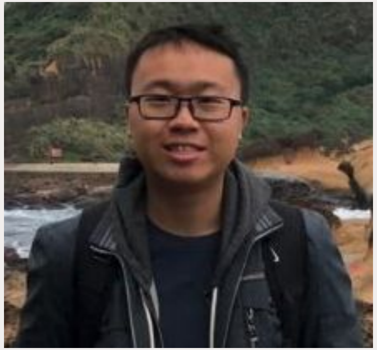
Sim Jing Wei, Chen Xiangzhen Samson, Tan Ding Xiang

Team Members

Wednesday 04/08/2021 Newsletter Edition: 001

Jing Wei

Samson



Ding Xiang

Table of Contents

Wednesday 15/09/2021

Newsletter

Edition: 002

01 **New Changes**

What we have change

02 **Project Progress**

Our current project stage

03 **Data Cleaning**

What we do

04 **Exploratory Data Analysis**

Showcase our insights and findings

05 **Our Next Stage**

What we are going to do next

06 **Reflection**

What we have learn so far

“Fake News Detector on COVID-19”

Wednesday 15/09/2021 Newsletter Edition: 002

01

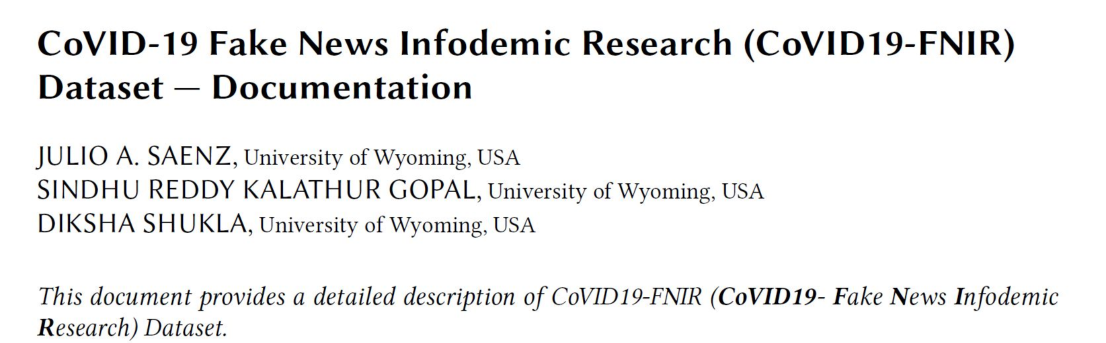
New Changes

New Changes & Focus

Wednesday 15/09/2021 Newsletter Edition: 002

Focus on COVID-19 Fake News

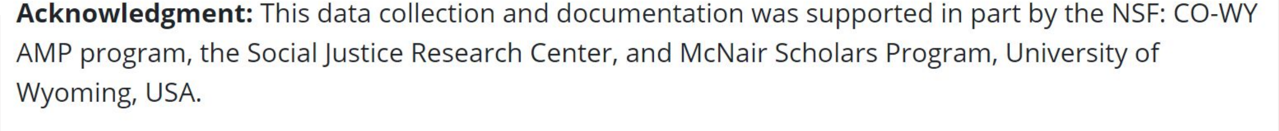
- Prevalence of COVID-19
- Spread of COVID-19 related fake news and information
- Focus on detecting fake COVID-19 news



CoVID-19 Fake News Infodemic Research (CoVID19-FNIR) Dataset — Documentation

JULIO A. SAENZ, University of Wyoming, USA
SINDHU REDDY KALATHUR GOPAL, University of Wyoming, USA
DIKSHA SHUKLA, University of Wyoming, USA

*This document provides a detailed description of CoVID19-FNIR (**CoVID19**- **F**ake **N**ews **I**nfodemic **R**esearch) Dataset.*

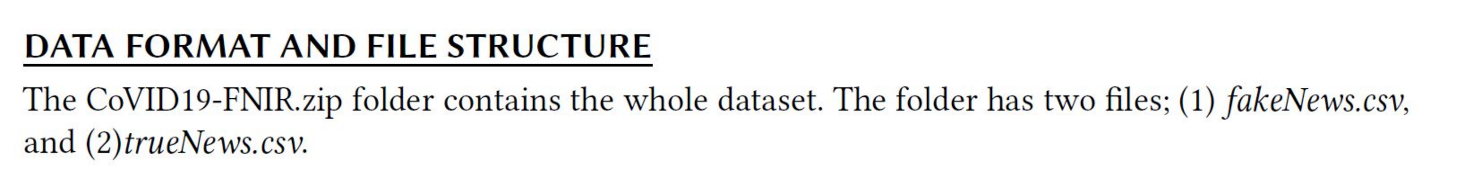


Acknowledgment: This data collection and documentation was supported in part by the NSF: CO-WY AMP program, the Social Justice Research Center, and McNair Scholars Program, University of Wyoming, USA.

New Changes & Focus

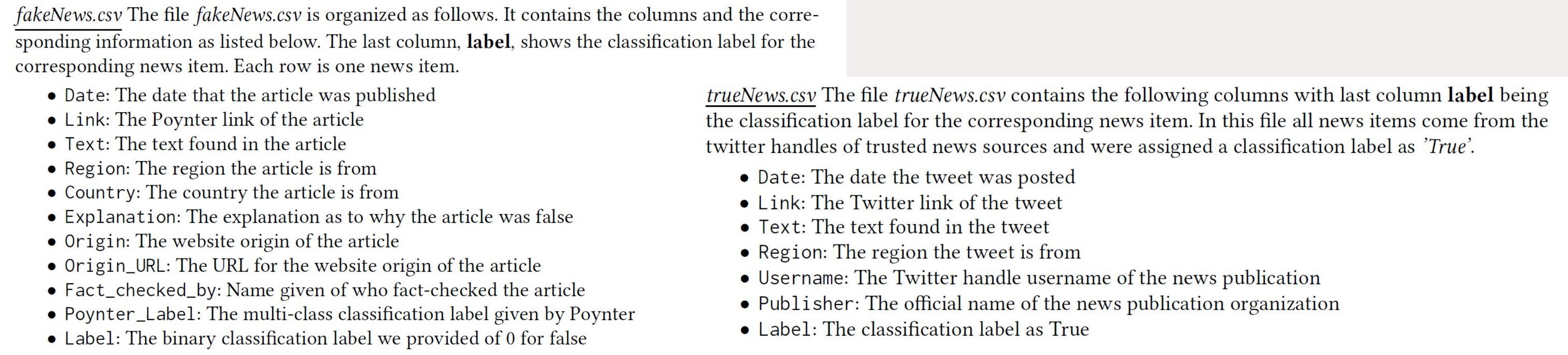
Wednesday 15/09/2021 Newsletter Edition: 002

New Datasets (COVID-19 Related)



DATA FORMAT AND FILE STRUCTURE

The CoVID19-FNIR.zip folder contains the whole dataset. The folder has two files; (1) *fakeNews.csv*, and (2)*trueNews.csv*.



fakeNews.csv The file *fakeNews.csv* is organized as follows. It contains the columns and the corresponding information as listed below. The last column, **label**, shows the classification label for the corresponding news item. Each row is one news item.

- `Date`: The date that the article was published
- `Link`: The Poynter link of the article
- `Text`: The text found in the article
- `Region`: The region the article is from
- `Country`: The country the article is from
- `Explanation`: The explanation as to why the article was false
- `Origin`: The website origin of the article
- `Origin_URL`: The URL for the website origin of the article
- `Fact_checked_by`: Name given of who fact-checked the article
- `Poynter_Label`: The multi-class classification label given by Poynter
- `Label`: The binary classification label we provided of 0 for false

trueNews.csv The file *trueNews.csv* contains the following columns with last column **label** being the classification label for the corresponding news item. In this file all news items come from the twitter handles of trusted news sources and were assigned a classification label as *'True'*.

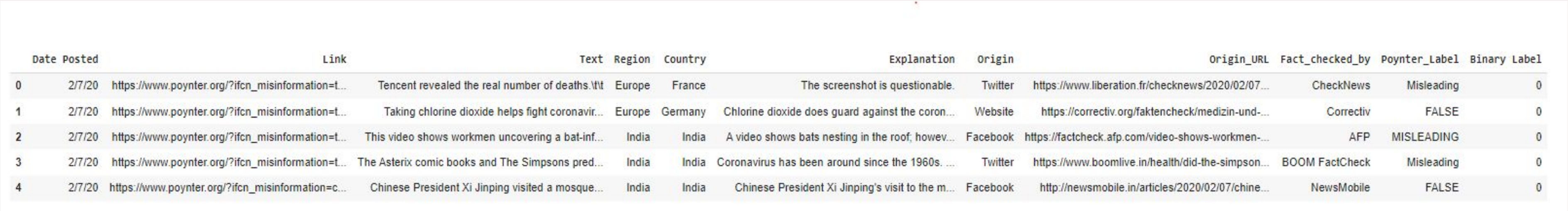
- `Date`: The date the tweet was posted
- `Link`: The Twitter link of the tweet
- `Text`: The text found in the tweet
- `Region`: The region the tweet is from
- `Username`: The Twitter handle username of the news publication
- `Publisher`: The official name of the news publication organization
- `Label`: The classification label as True

Source: https://ieee-dataport.org/open-access/covid-19-fake-news-infodemic-research-dataset-covid19-fnir-dataset

New Changes

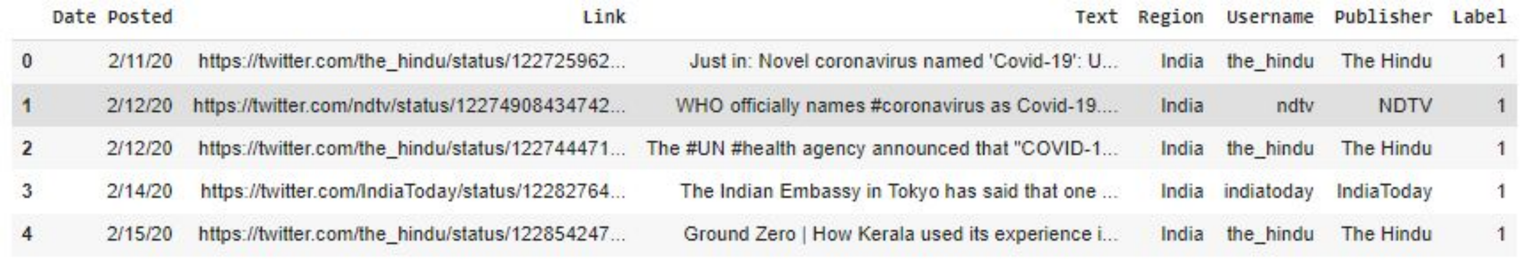
Wednesday 15/09/2021 Newsletter Edition: 002

Fake News Dataset



	Date Posted	Link	Text	Region	Country	Explanation	Origin	Origin_URL	Fact_checked_by	Poynter_Label	Binary Label
0	2/7/20	https://www.poynter.org/?ifcn_misinformation=t...	Tencent revealed the real number of deaths.\t\t	Europe	France	The screenshot is questionable.	Twitter	https://www.liberation.fr/checknews/2020/02/07...	CheckNews	Misleading	0
1	2/7/20	https://www.poynter.org/?ifcn_misinformation=t...	Taking chlorine dioxide helps fight coronavir...	Europe	Germany	Chlorine dioxide does guard against the coron...	Website	https://correctiv.org/faktencheck/medizin-und-...	Correctiv	FALSE	0
2	2/7/20	https://www.poynter.org/?ifcn_misinformation=t...	This video shows workmen uncovering a bat-inf...	India	India	A video shows bats nesting in the roof; howev...	Facebook	https://factcheck.afp.com/video-shows-workmen-...	AFP	MISLEADING	0
3	2/7/20	https://www.poynter.org/?ifcn_misinformation=t...	The Asterix comic books and The Simpsons pred...	India	India	Coronavirus has been around since the 1960s. ...	Twitter	https://www.boomlive.in/health/did-the-simpson...	BOOM FactCheck	Misleading	0
4	2/7/20	https://www.poynter.org/?ifcn_misinformation=c...	Chinese President Xi Jinping visited a mosque...	India	India	Chinese President Xi Jinping's visit to the m...	Facebook	http://newsmobile.in/articles/2020/02/07/chine...	NewsMobile	FALSE	0

Real News Dataset



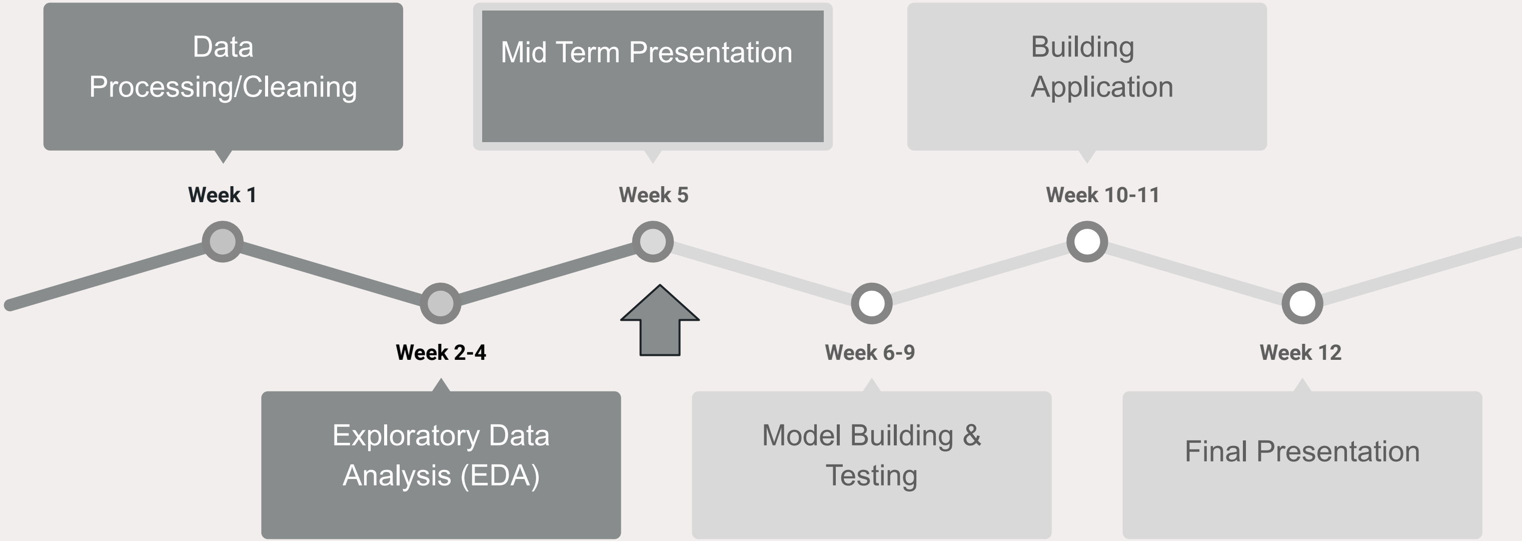
	Date Posted	Link	Text	Region	Username	Publisher	Label
0	2/11/20	https://twitter.com/the_hindu/status/122725962...	Just in: Novel coronavirus named 'Covid-19': U...	India	the_hindu	The Hindu	1
1	2/12/20	https://twitter.com/ndtv/status/12274908434742...	WHO officially names #coronavirus as Covid-19....	India	ndtv	NDTV	1
2	2/12/20	https://twitter.com/the_hindu/status/122744471...	The #UN #health agency announced that "COVID-1...	India	the_hindu	The Hindu	1
3	2/14/20	https://twitter.com/IndiaToday/status/12282764...	The Indian Embassy in Tokyo has said that one ...	India	indiatoday	IndiaToday	1
4	2/15/20	https://twitter.com/the_hindu/status/122854247...	Ground Zero | How Kerala used its experience i...	India	the_hindu	The Hindu	1

02

Project Progress

Project Progress

Wednesday 15/09/2021 Newsletter Edition: 002



Data Processing/Cleaning

Week 1

Exploratory Data Analysis (EDA)

Week 2-4

Mid Term Presentation

Week 5

Model Building & Testing

Week 6-9

Building Application

Week 10-11

Final Presentation

Week 12

03 Data Cleaning

Data Cleaning

Wednesday 04/08/2021 Newsletter Edition: 001

Removal of Unnecessary Text

- URL Links
- Usernames e.g. @users
- Non-Alphanumeric Characters
- Stop Words

Removal of Duplicates

- There were a total 322 Duplicates News

Removal of Unnecessary Columns

- To be able to merge both the Real & Fake Datasets

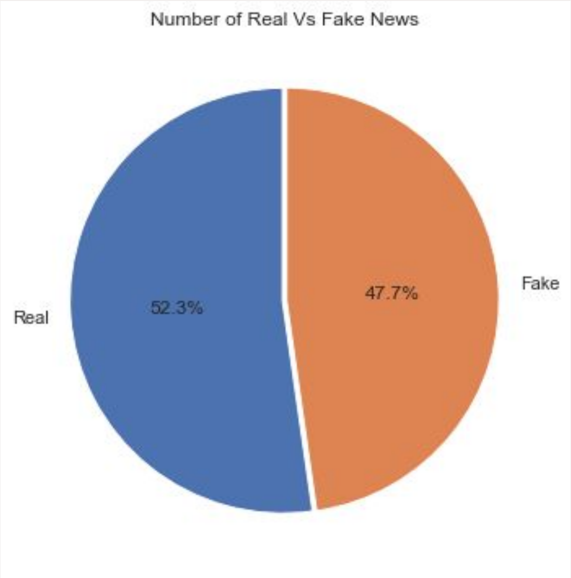
04

Exploratory Data Analysis

Insights

Wednesday 04/08/2021 | Newsletter | Edition: 001

Dataset

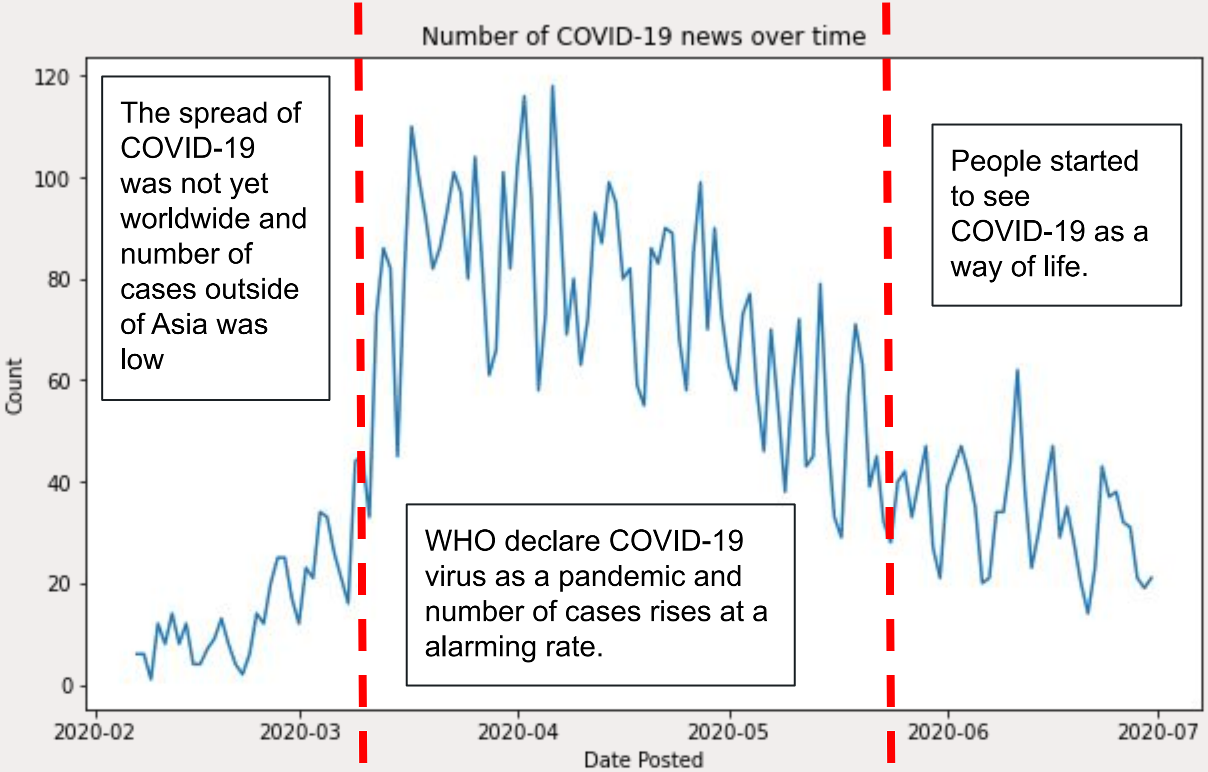


- Total of 7256 COVID-19 News/Tweets
- 3792 of them are Real, while 3464 of them are Fake

Insights

Wednesday 04/08/2021 | Newsletter | Edition: 001

Trends



The spread of COVID-19 was not yet worldwide and number of cases outside of Asia was low

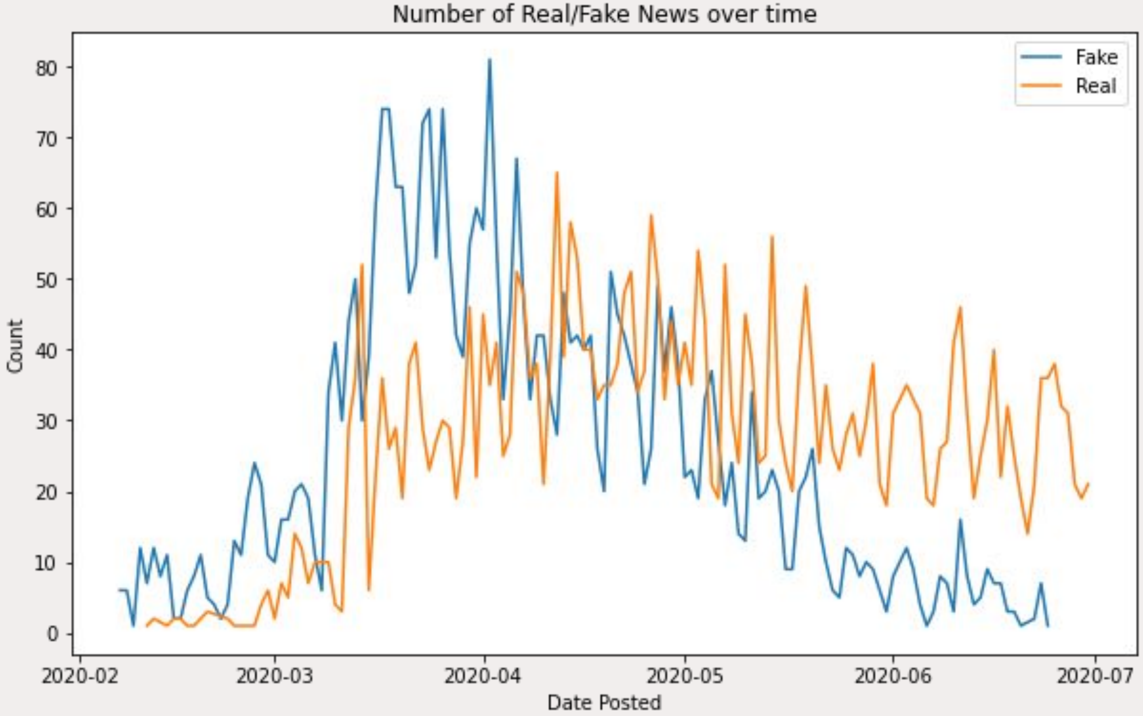
WHO declare COVID-19 virus as a pandemic and number of cases rises at a alarming rate.

People started to see COVID-19 as a way of life.

Insights

Wednesday 04/08/2021 Newsletter Edition: 001

Trends



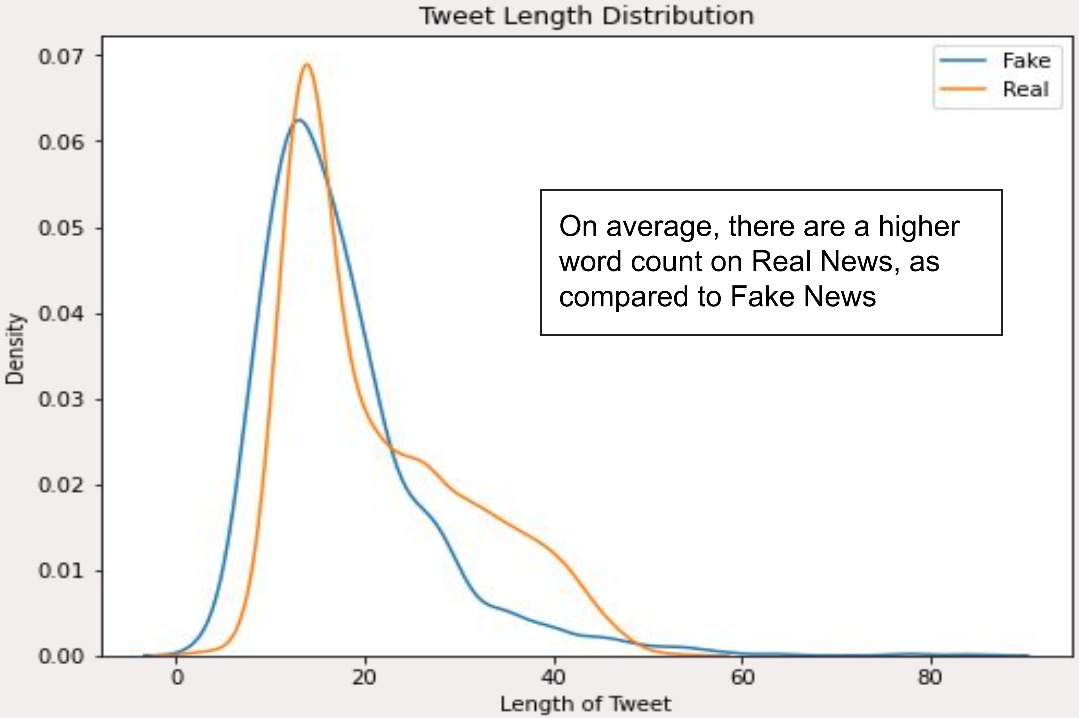
- The number of Fake News peaked during the early phrase of the pandemic, where people are new and unfamiliar with the new virus. Eventually, the number started to decrease over time.

- The number of Real News have increased over time, eventually more than the number of Fake News. This is perhaps due to more researches and studies being done on the virus itself.

Insights

Wednesday 04/08/2021 Newsletter Edition: 001

Tweet Length (Word Count)

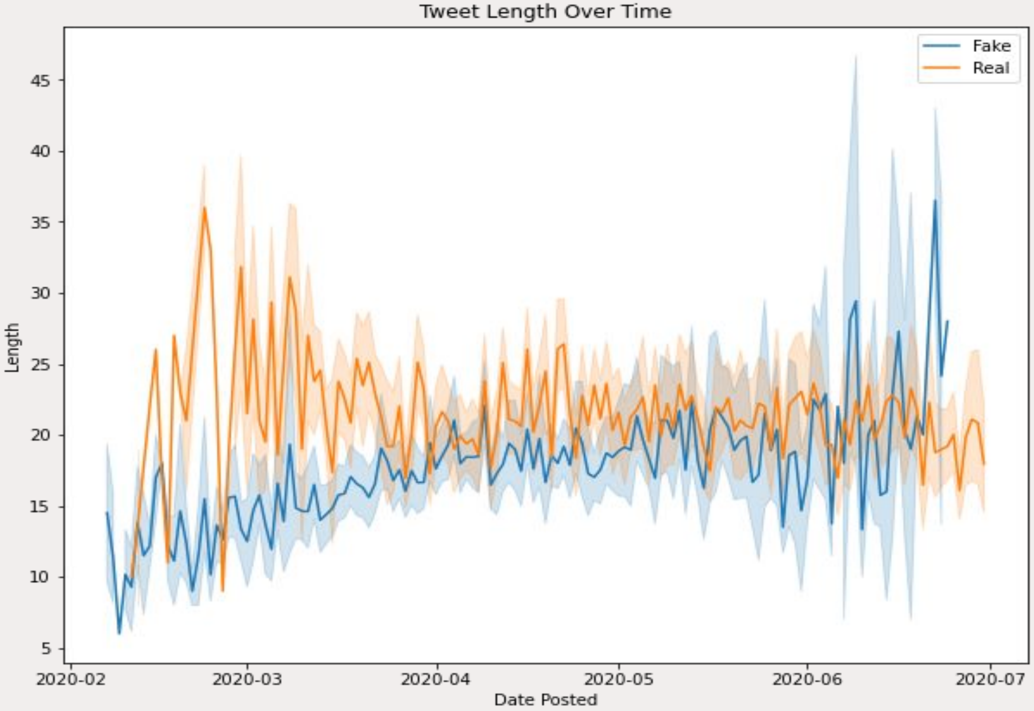


On average, there are a higher word count on Real News, as compared to Fake News

Insights

Wednesday 04/08/2021 Newsletter Edition: 001

Tweets Length (Word Count)



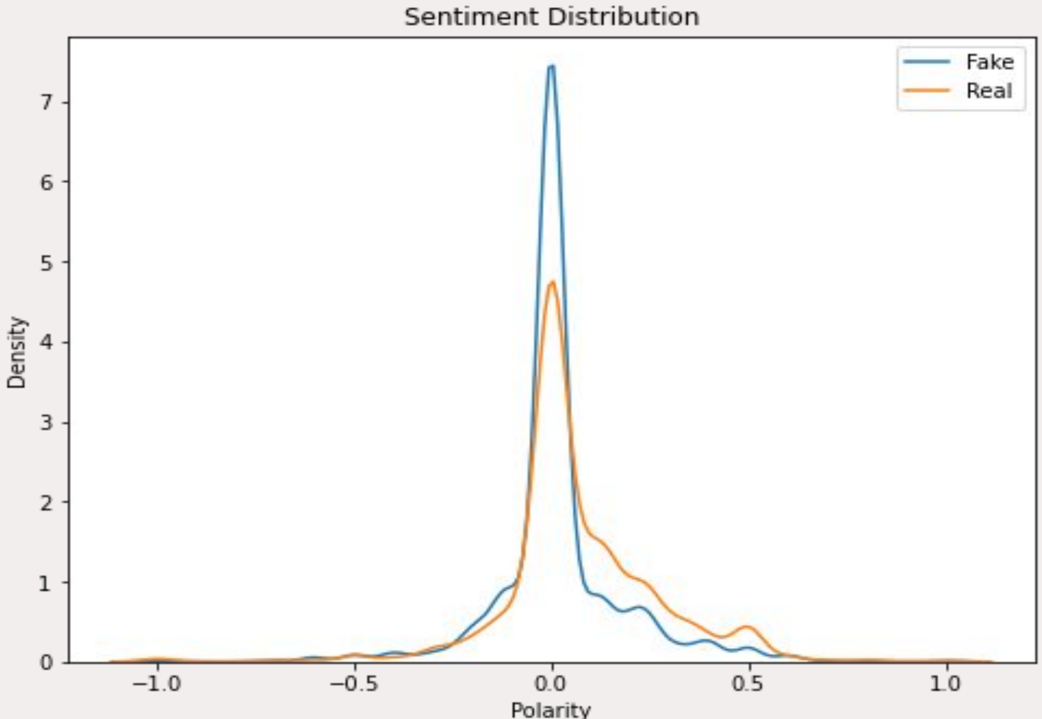
- For Fake News, the length of tweets remains roughly the same over time, despite some large fluctuation at the start.

- For Real News, the length of tweets increases over time. This is perhaps due to having more information about COVID-19, which lead to more content.

Insights

Wednesday 04/08/2021 Newsletter Edition: 001

Sentiment Analysis (Polarity)

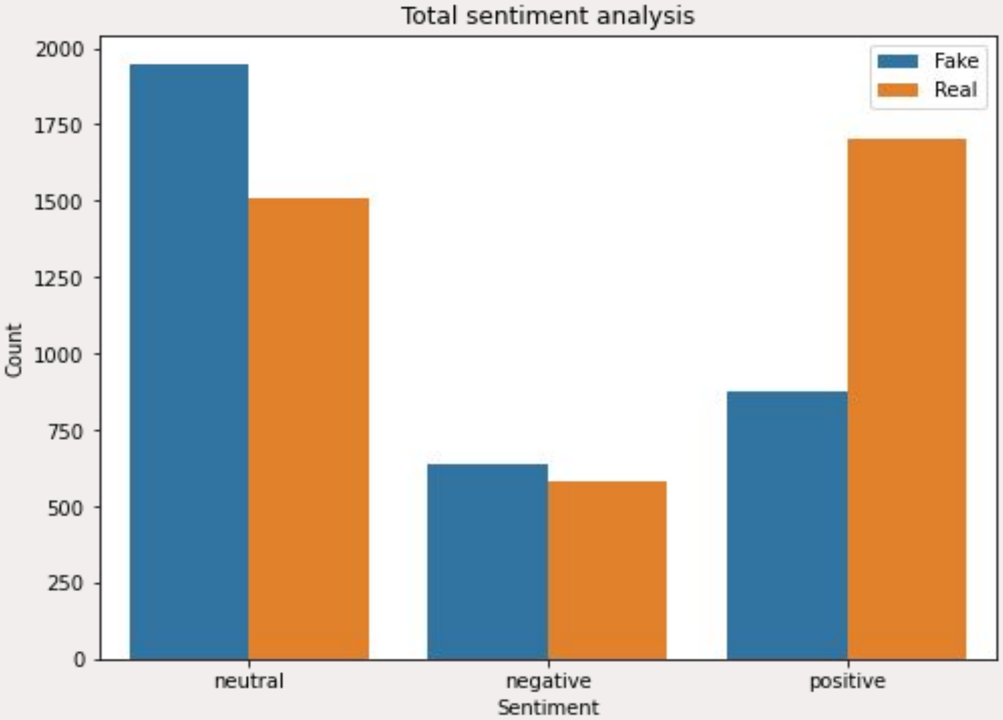


- Polarity is used to measure the sentiment or emotion of the text.
- The closer the value to 1, the more positive the text is, and the closer the value to -1, the more negative the text is.
- In general, majority of the tweets have a polarity score of around 0 (netural). But, Real News tend to be more positive, as compared to Fake News.

Insights

Wednesday 04/08/2021 Newsletter Edition: 001

Sentiment Analysis (Polarity)

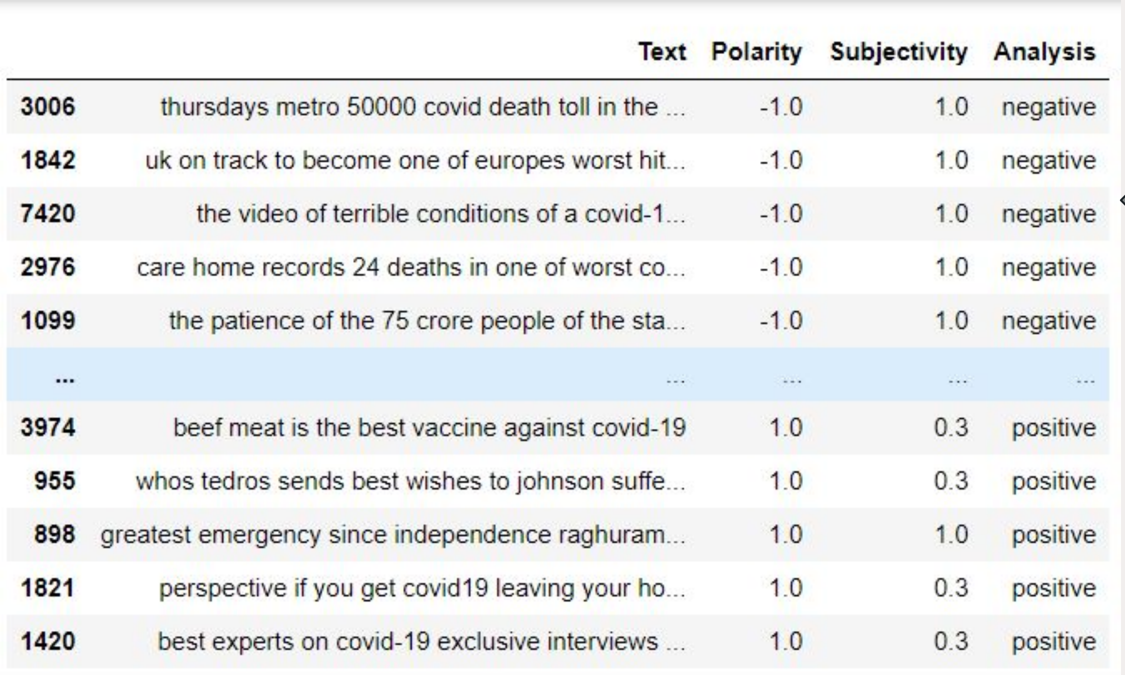


Real News tend to be more positive, while Fake News tend to be slightly more negative

Insights

Wednesday 04/08/2021 Newsletter Edition: 001

Sentiment Analysis (Polarity)



	Text	Polarity	Subjectivity	Analysis
3006	thursdays metro 50000 covid death toll in the ...	-1.0	1.0	negative
1842	uk on track to become one of europes worst hit...	-1.0	1.0	negative
7420	the video of terrible conditions of a covid-1...	-1.0	1.0	negative
2976	care home records 24 deaths in one of worst co...	-1.0	1.0	negative
1099	the patience of the 75 crore people of the sta...	-1.0	1.0	negative
...	...	...	...	...
3974	beef meat is the best vaccine against covid-19	1.0	0.3	positive
955	whos tedros sends best wishes to johnson suffe...	1.0	0.3	positive
898	greatest emergency since independence raghuram...	1.0	1.0	positive
1821	perspective if you get covid19 leaving your ho...	1.0	0.3	positive
1420	best experts on covid-19 exclusive interviews ...	1.0	0.3	positive

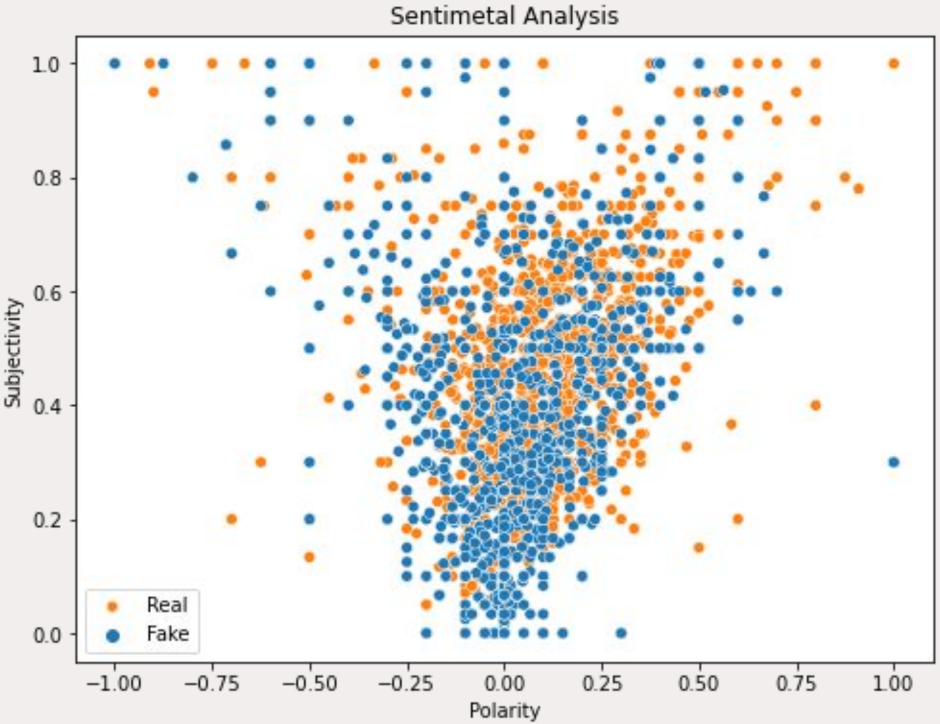
Top 5 Most Negative Tweets

Top 5 Most Positive Tweets

Insights

Wednesday 04/08/2021 Newsletter Edition: 001

Sentiment Analysis (Polarity VS Subjectivity)



- Subjectivity refers to how subjective or objective the text is.

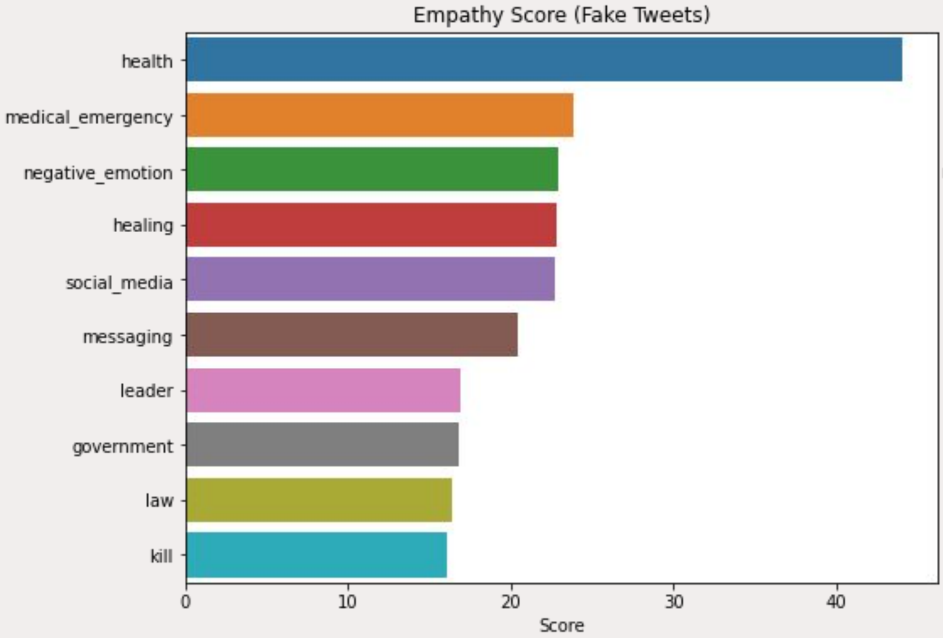
- The closer the value to 1, the more subjective the text is, and the closer the value to 0, the more objective the text is.

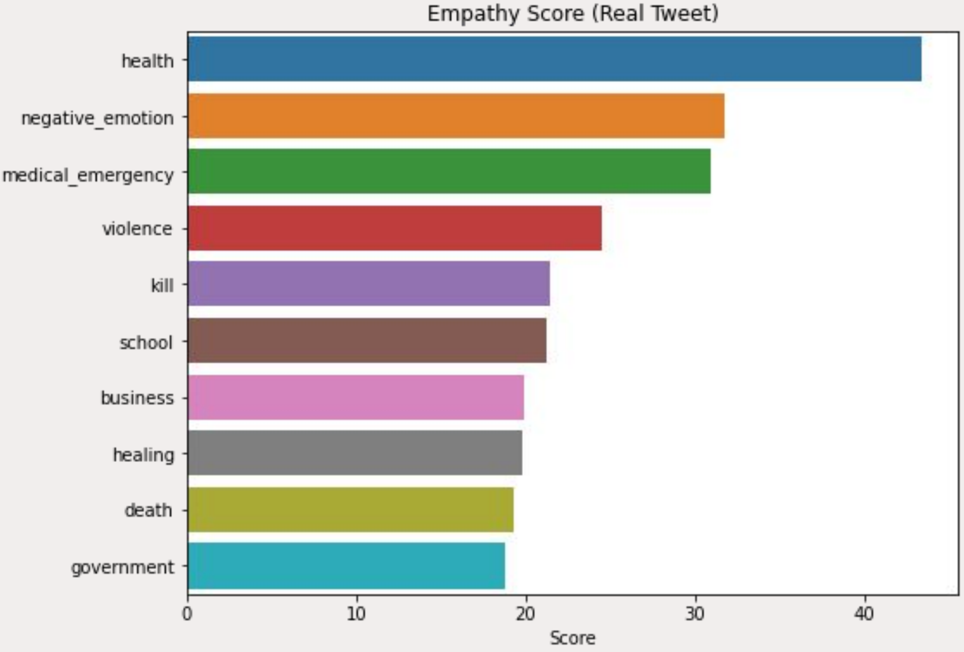
- In general, the more negative/positive (polarity) the tweet is, it will also be more subjective.

Insights

Wednesday 04/08/2021 Newsletter Edition: 001

Empathy Analysis


Empathy Score (Fake Tweets)

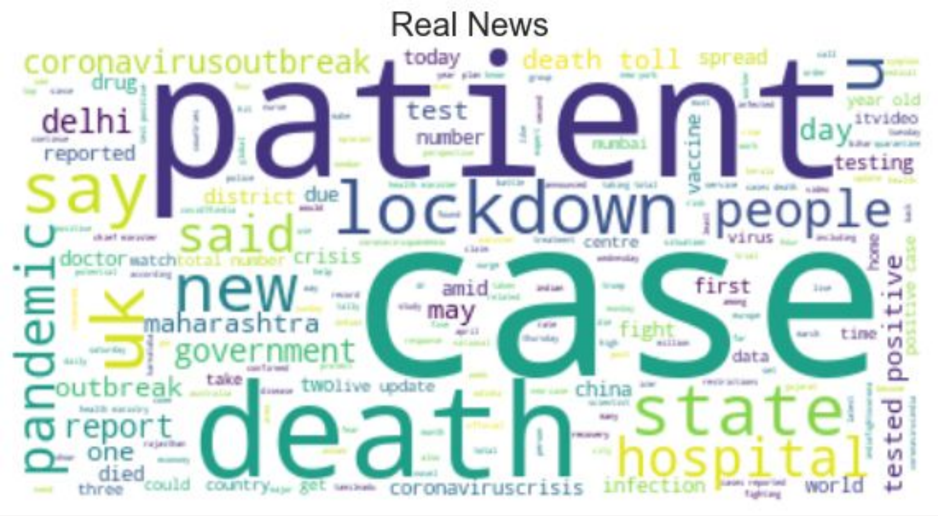

Empathy Score (Real Tweet)

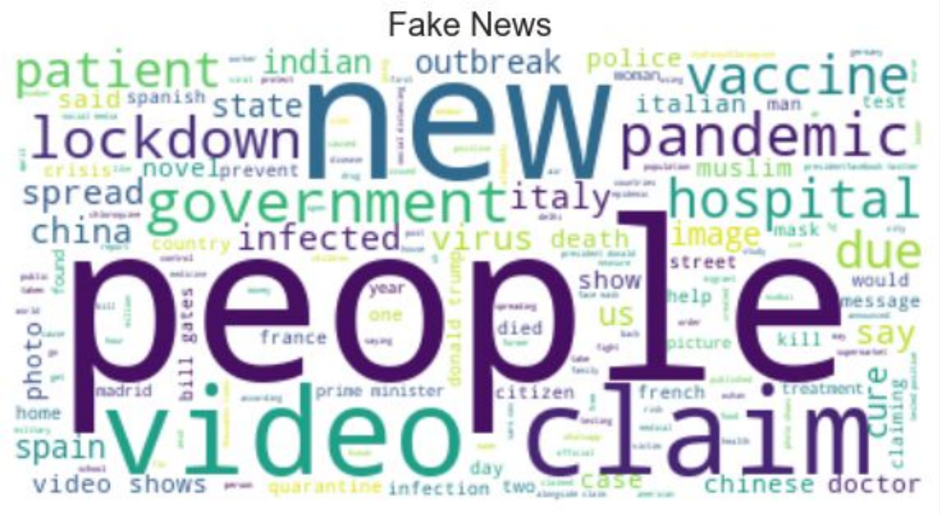
Empathy Analysis uses words/terms to measure the emotion/feeling of the text

Insights

Wednesday 04/08/2021 Newsletter Edition: 001

Word Cloud


Real News


Fake News

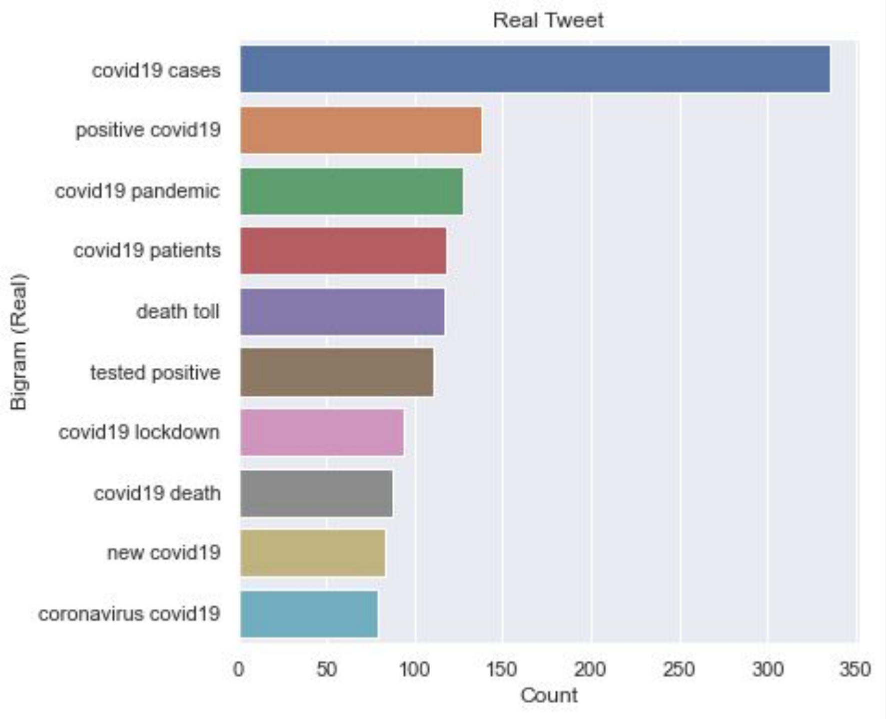
Real News focuses more on reporting the number of COVID 19 cases and death toll.

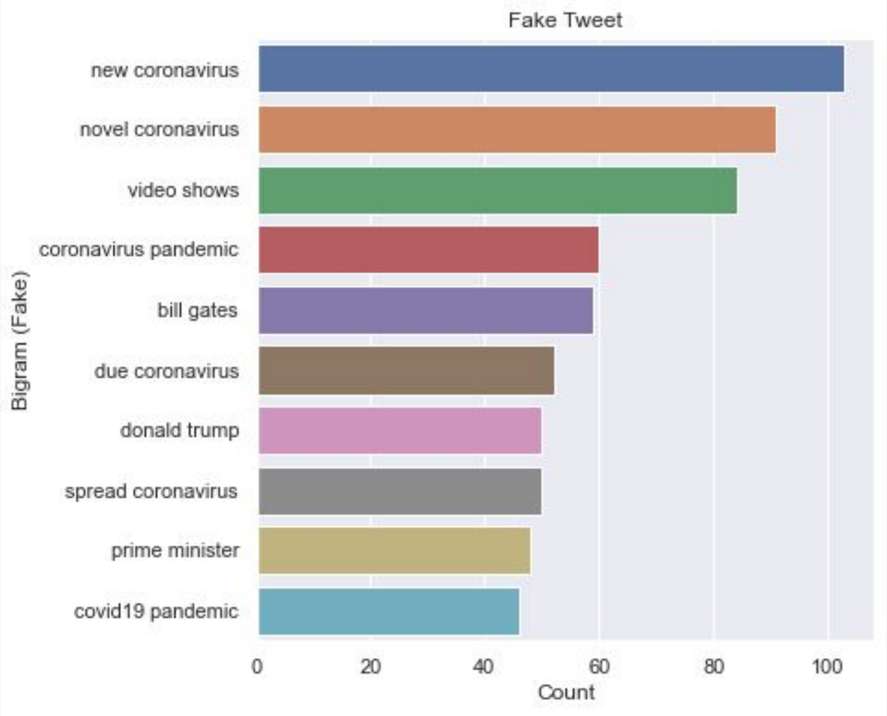
Fake News focuses more on spreading rumors through video.

Insights

Wednesday 04/08/2021 Newsletter Edition: 001

Bigram (Top 10)

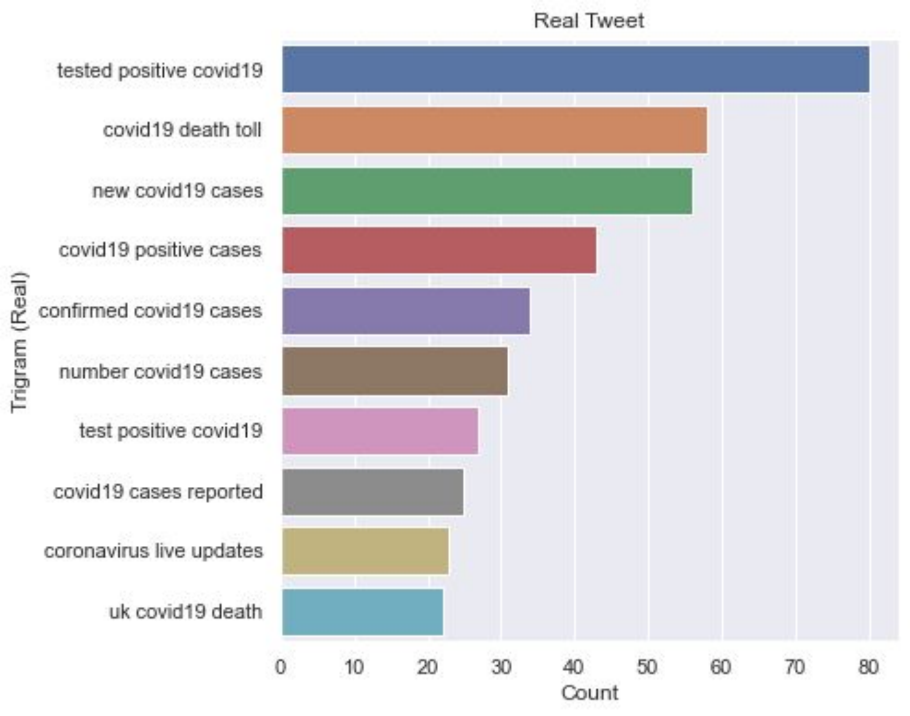


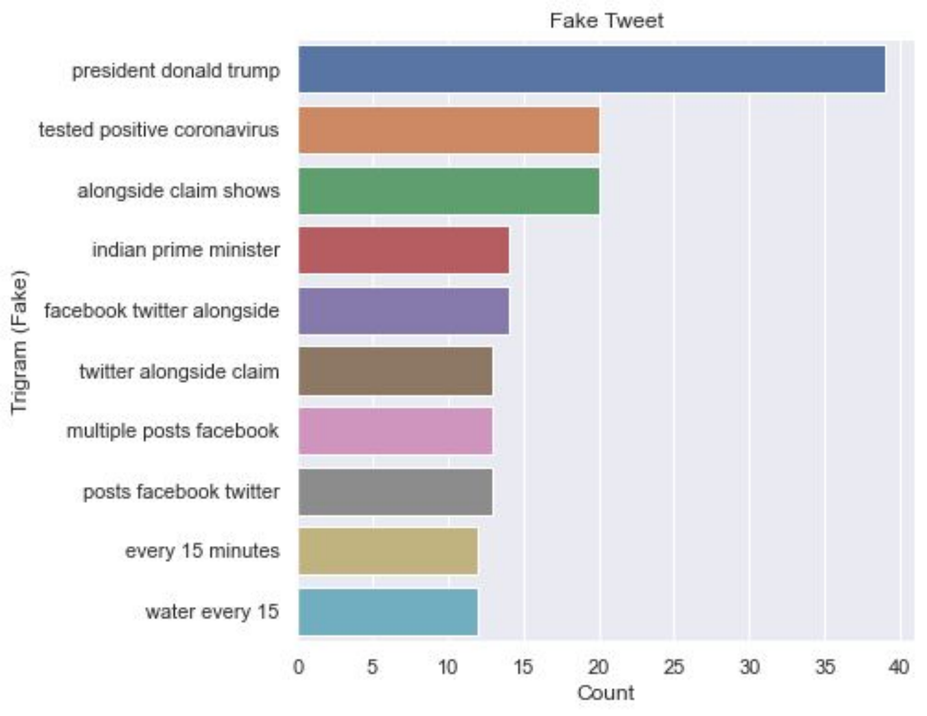


Insights

Wednesday 04/08/2021 Newsletter Edition: 001

Trigram (Top 10)

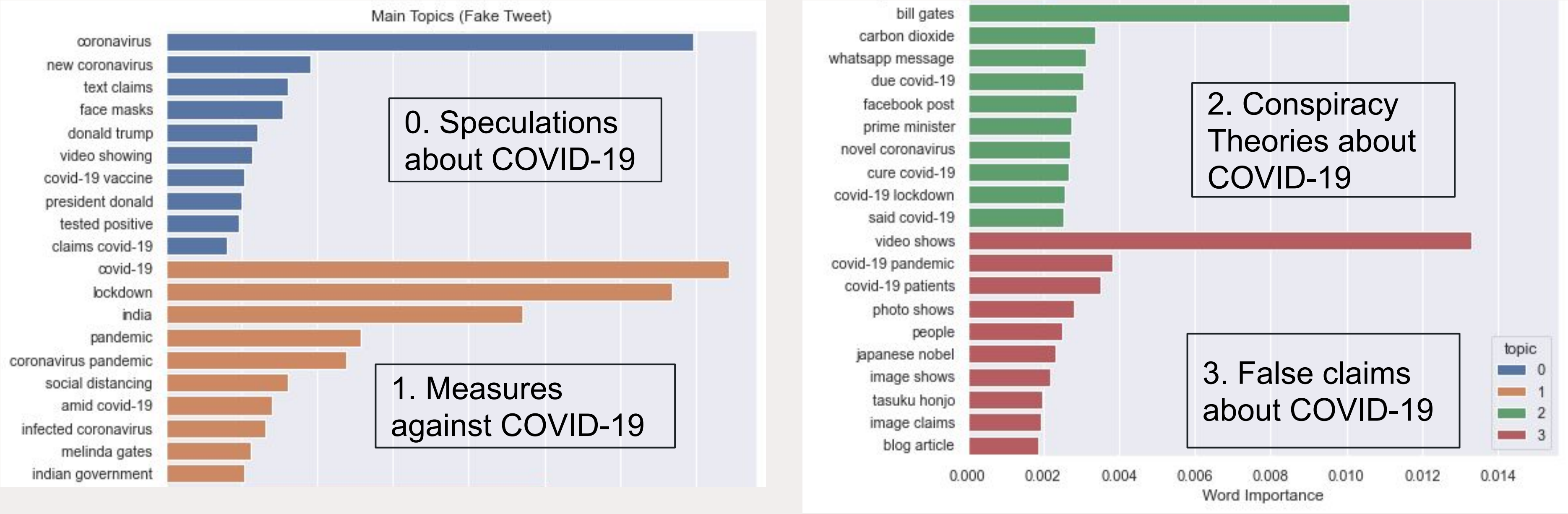




Insights

Wednesday 04/08/2021 Newsletter Edition: 001

Topic Modelling for Fake News



Main Topics (Fake Tweet)

0. Speculations about COVID-19

1. Measures against COVID-19

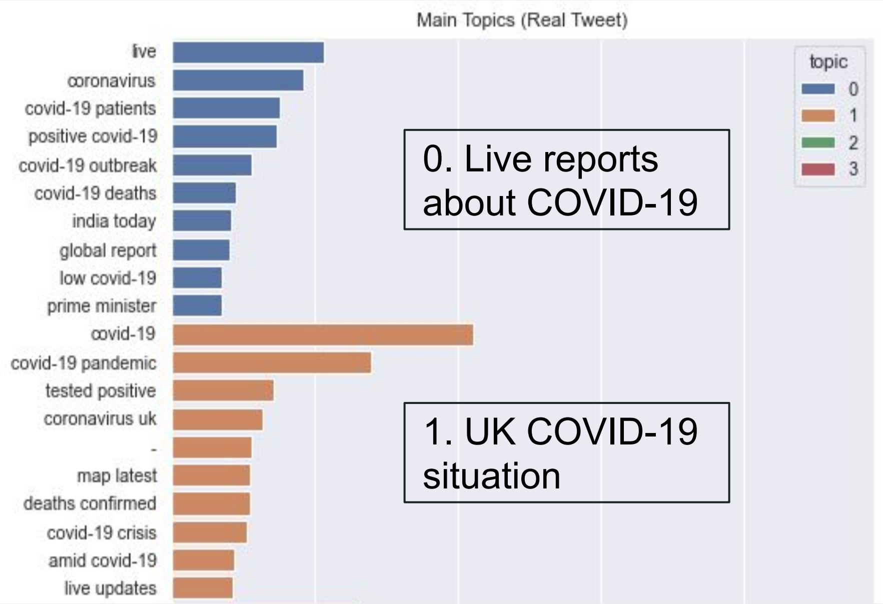
2. Conspiracy Theories about COVID-19

3. False claims about COVID-19

Insights

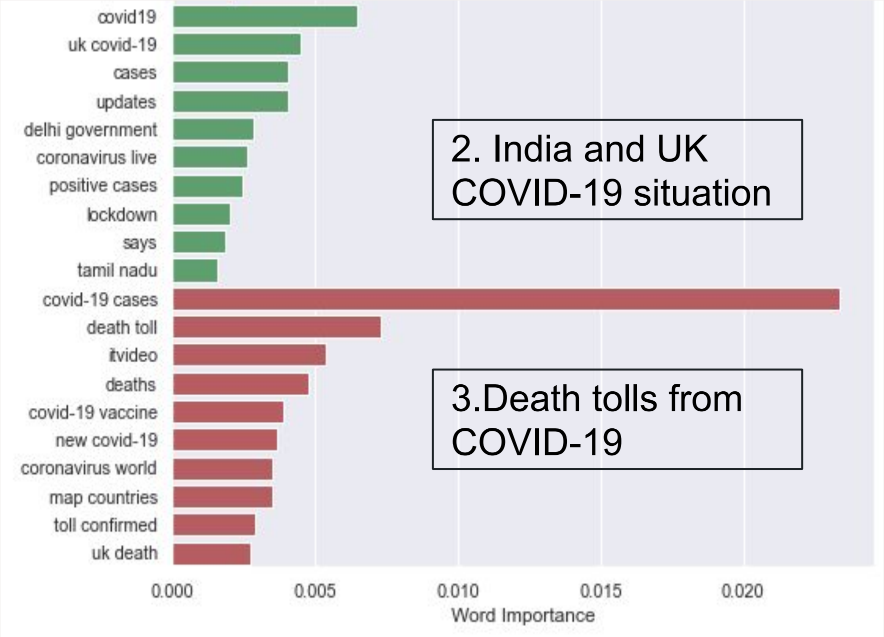
Wednesday 04/08/2021 | Newsletter | Edition: 001

Topic Modelling for Real News



0. Live reports about COVID-19

1. UK COVID-19 situation



2. India and UK COVID-19 situation

3.Death tolls from COVID-19

Feature Engineering

Wednesday 04/08/2021 Newsletter Edition: 001

Feature Engineering

- Length of Tweets
- Word Count of each Tweet
- Sentence Count of each Tweet
- Number of Capital Characters
- Number of Punctuations
- Number of Unique Words
- Stopwords Count
- Average Word Length
- Average Sentence Length

05

Our Next Stage

Our Next Stage

Wednesday 04/08/2021 Newsletter Edition: 001

Feature Selection

- Dimensionality Reduction

Tokenization

- Lemmatization
- Word2Vec with n-gram

Our Next Stage

Wednesday 04/08/2021 Newsletter Edition: 001

Model Training & Testing

- Support Vector Machine (SVM)
- Random Forest (RF)
- LSTM
- BERT

Evaluation Metric & Model Selection

- Accuracy
- F-Score (Precision & Recall)
- K-fold Cross Validation

Development & Deployment

- Telegram Bot / Web Application
- Web Scraping
- Heroku

06

Reflection

Difficulties faced

Wednesday 04/08/2021 Newsletter Edition: 001

Difficulties

- Difficulty in finding good datasets
- New to different NLP techniques
- Much discussion in deciding which features to include during text cleaning

Thank You! :)

Wednesday 04/08/2021 Newsletter Edition: 001

THE SUNDAY JOURNAL
Trump's 'Merry Ch
President wants to take political correctness out of the national holiday
Joy to the weed! Legal marijuana lands on m
Homegrown options increasingly popular
KOHLER