UNIVERSITÄT DUISBURG ESSEN

Open-Minded

OTTO-FRIEDRICH-UNIVERSITÄT BAMBERG UNIVERSITY OF BAMBERG

FakeCovid- A Multilingual Cross-domain Dataset for COVID-19

Gautam Kishore Shahi, Durgesh Nandini

University of Duisburg-Essen, Germany, University of Bamberg Germany

21st September 2020

Contents

- Introduction
- Our Contribution
- Datasets
- Class Normalization
- Dataset I & II
- Open Questions
- References

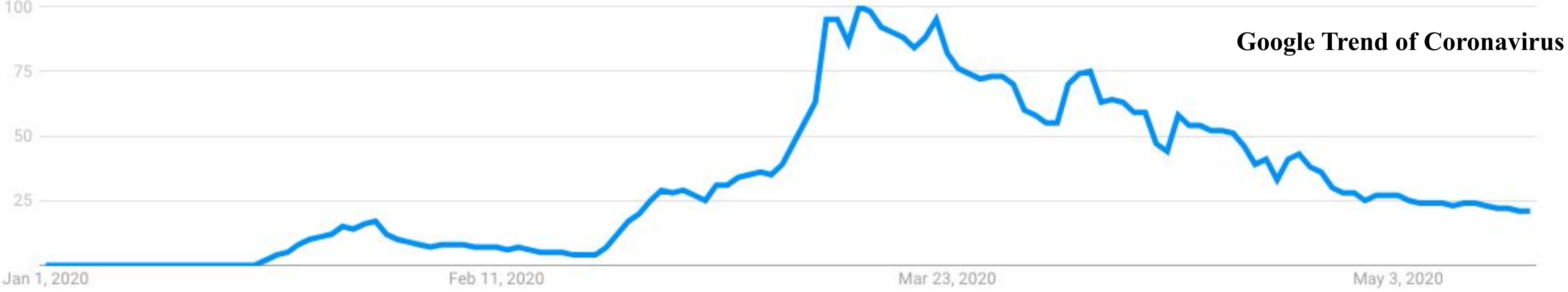
Introduction

- ***Infodemic*** makes difficult for users to find reliable sources for any claim made on the pandemic, either on the news or social media.

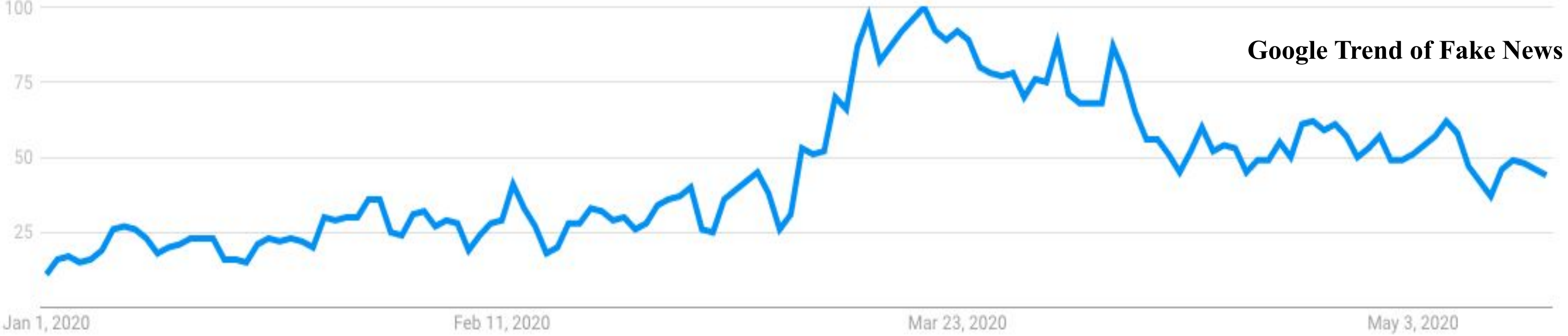
- **Lack of corpus to test methods for fake news detection on the pandemic.**

- More than **92** fact-checking websites in **40** languages are working around the globe.

Google Trend on Coronavirus and Fake News



Google Trend of Coronavirus



Google Trend of Fake News

Our Contribution

- **FakeCovid**: A very timely study on misinformation on the COVID-19 pandemic. Answering where COVID-19 misinformation originates from, and how it spreads.

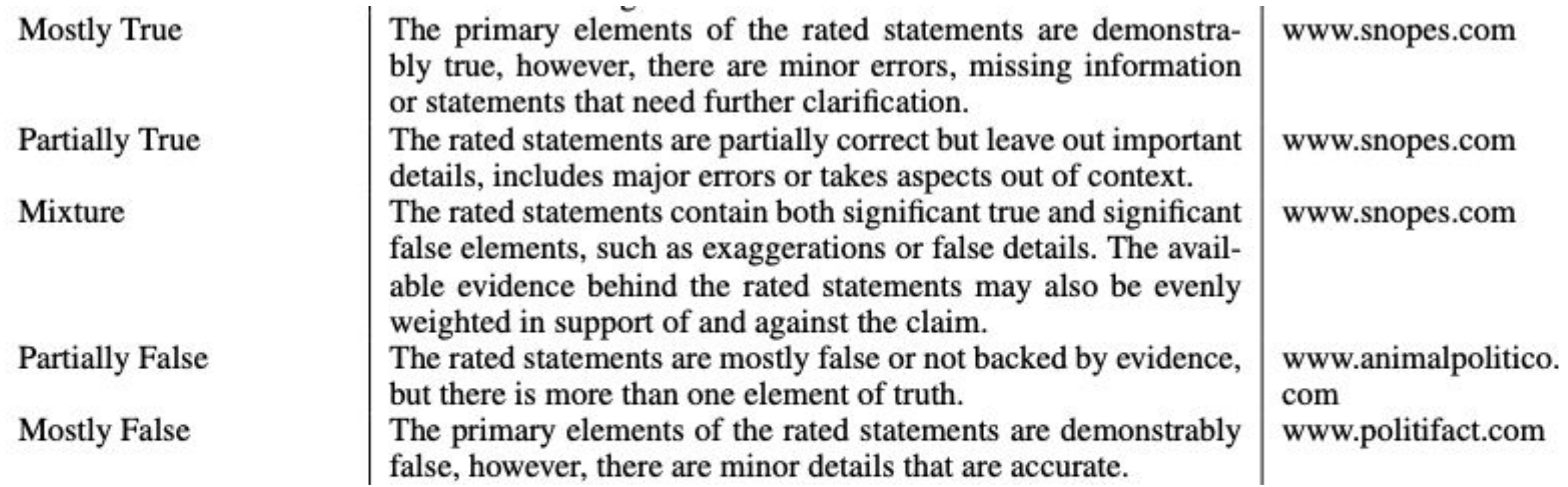
- Explaining what makes COVID-19 misinformation distinct from other tweets on COVID-19.

- Merged **86** classes into **4** classes of fact checked articles.

Datasets

- We used two different datasets, one is the fact checked articles and another is sampled tweet from twitter.
    - **Dataset I** - 7623 news articles **Dataset II** - 1500 labelled tweets and 164805 random sampled tweet
- **Dataset I-** Took Poynter and Snopes as a reference data hub. Poynter has initiated International Fact-Checking Network (IFCN) to bring fact-checkers worldwide to a single platform. IFCN maintains a separate data hub for COVID-19. IFCN uses hashtags #CoronaVirusFacts, #DatosCoronaVirus.
- **Dataset II -** It was crawled using finding the link of the embedded tweets from the fact checked articles. Random sampled tweets collected using different hashtags like #covid19,#coronavirus.

Class Used by IFCN



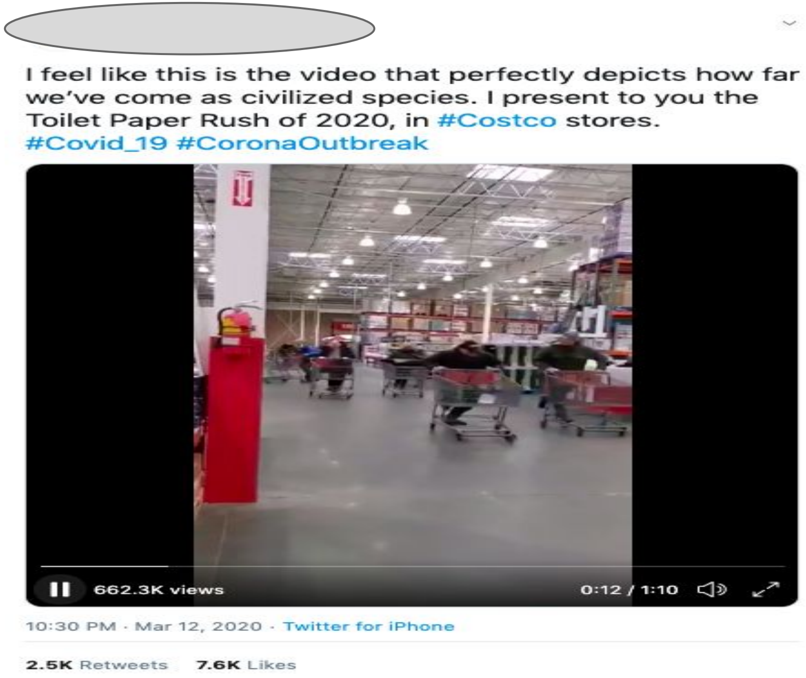
Mostly True	The primary elements of the rated statements are demonstrably true, however, there are minor errors, missing information or statements that need further clarification.	www.snopes.com
Partially True	The rated statements are partially correct but leave out important details, includes major errors or takes aspects out of context.	www.snopes.com
Mixture	The rated statements contain both significant true and significant false elements, such as exaggerations or false details. The available evidence behind the rated statements may also be evenly weighted in support of and against the claim.	www.snopes.com
Partially False	The rated statements are mostly false or not backed by evidence, but there is more than one element of truth.	www.animalpolitico.com
Mostly False	The primary elements of the rated statements are demonstrably false, however, there are minor details that are accurate.	www.politifact.com

Class Normalisation

We normalised verdicts by manually mapping them to a score of 1 to 4 (1='False', 2='Partially False', 3='True', 4='Others') based on the definitions provided by the fact-checking organisations. Our definition for the four categories are as follows:

- **False**: Claims of an article are untrue.
- **Partially False**: Claims of an article are a mixture of true and false information. The article contains partially true and partially false information, but it can not be considered as 100% true. It includes articles of type, partially false, partially true, mostly true, miscaptioned, misleading etc.
- **True**: This rating indicates that the primary elements of a claim are demonstrably true.
- **Other**: An article that cannot be categorised as true, false or partially false due to lack of evidence about its claims. This category includes articles in dispute and unproven articles.

Examples of Misinformation

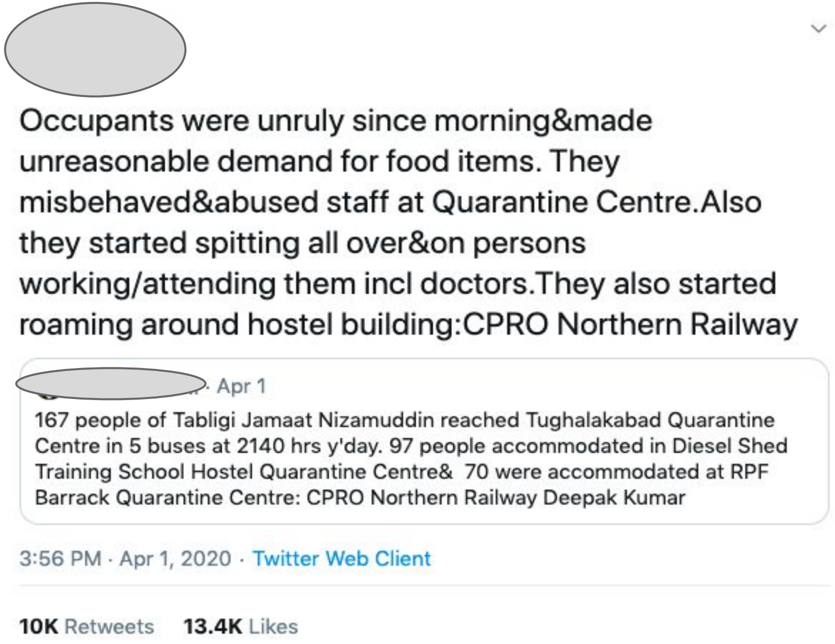


[redacted]

I feel like this is the video that perfectly depicts how far we've come as civilized species. I present to you the Toilet Paper Rush of 2020, in #Costco stores. #Covid_19 #CoronaOutbreak

10:30 PM · Mar 12, 2020 · Twitter for iPhone

2.5K Retweets 7.6K Likes



[redacted]

Occupants were unruly since morning&made unreasonable demand for food items. They misbehaved&abused staff at Quarantine Centre.Also they started spitting all over&on persons working/attending them incl doctors.They also started roaming around hostel building:CPRO Northern Railway

[redacted] · Apr 1

167 people of Tabligi Jamaat Nizamuddin reached Tughalakabad Quarantine Centre in 5 buses at 2140 hrs y'day. 97 people accommodated in Diesel Shed Training School Hostel Quarantine Centre& 70 were accommodated at RPF Barrack Quarantine Centre: CPRO Northern Railway Deepak Kumar

3:56 PM · Apr 1, 2020 · Twitter Web Client

10K Retweets 13.4K Likes

Examples of Misinformation of false and partially false tweet

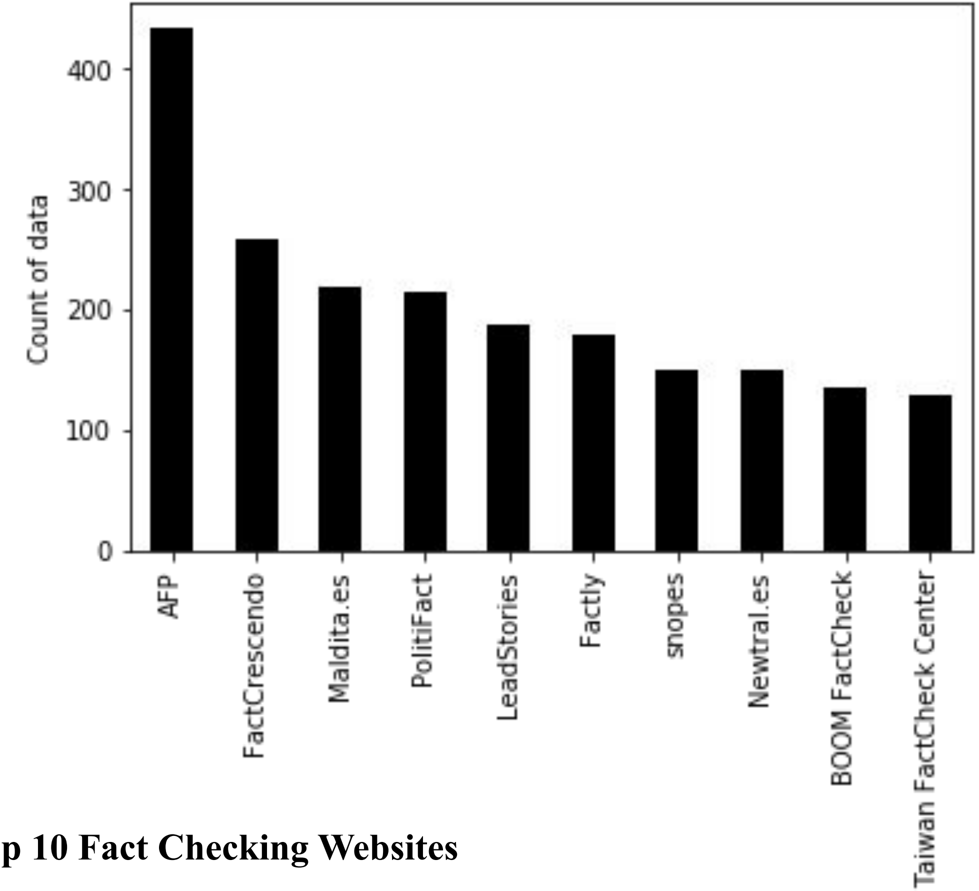
Dataset-I

- Crawled **7623** Fact check news article from **92** fact-checking websites in **40** languages, around **80%** of articles are of false category.
- Dataset contains fact check articles from **105** countries, top **10** countries (India, USA, Spain) contributes **65%** of the fake news.

Dataset-II

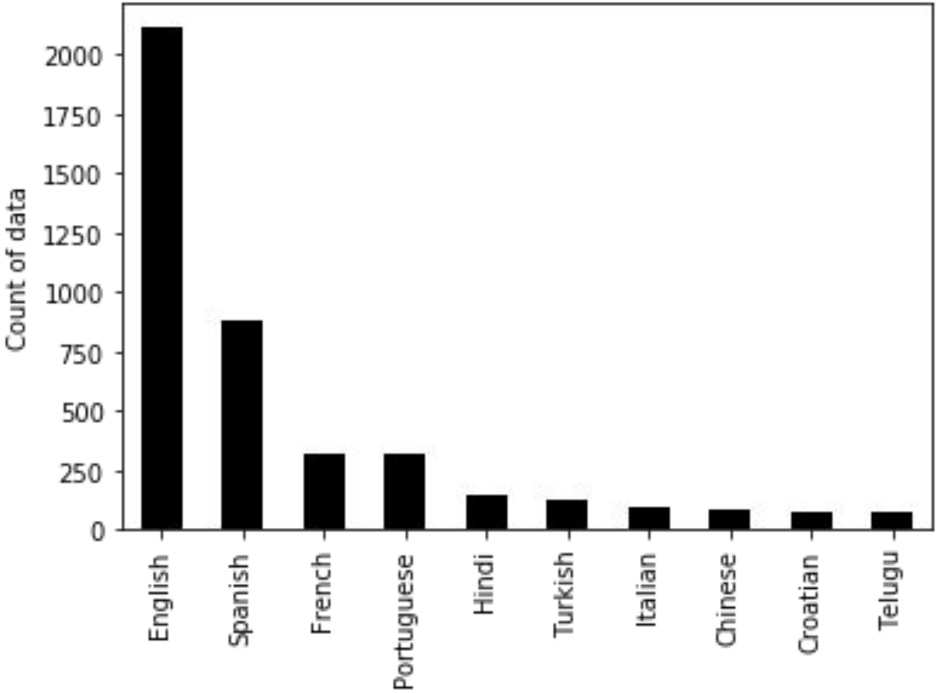
- Dataset II contains 1500 tweets which are classified as misinformation.
- These tweets are from From 1187 unique accounts from Twitter.
- These tweets are classified into four different classes into False, Partially False, True, Others.

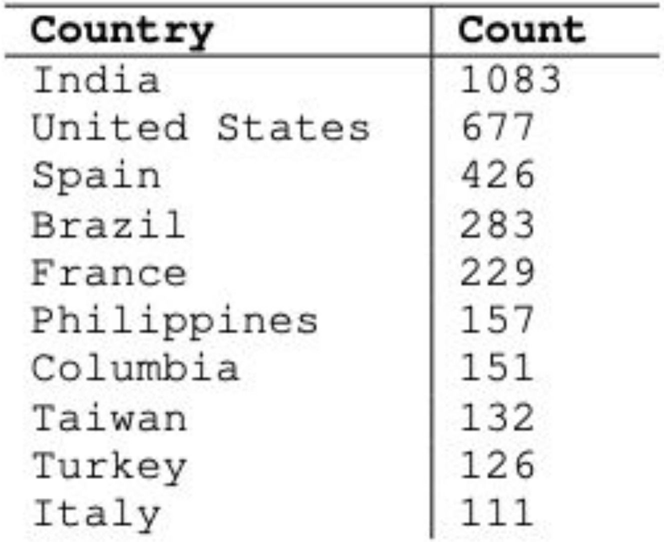
Dataset-I



Top 10 Fact Checking Websites

Dataset-I

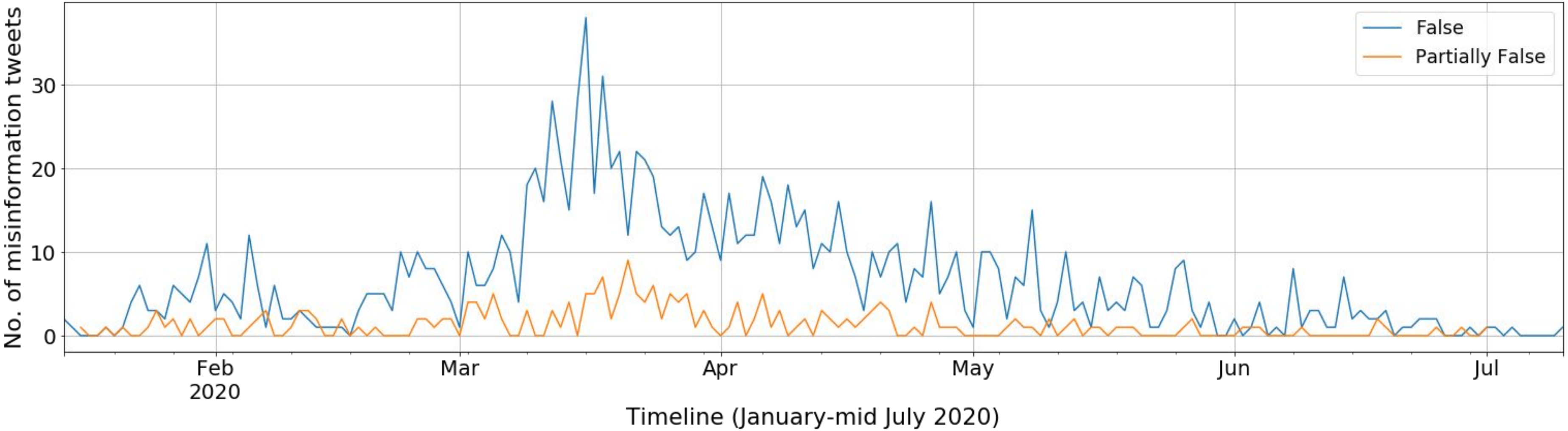




Country	Count
India	1083
United States	677
Spain	426
Brazil	283
France	229
Philippines	157
Columbia	151
Taiwan	132
Turkey	126
Italy	111

Top 10 Language of Fact checked news articles and Top 10 countries(based on 1st version)

Dataset-II



Timeline of misinformation tweets created during January 2020 to mid-July 2020

Open Questions

- Classification of Fake tweets in random sampled tweet using the labelled misinformation tweets.
- Finding trends/similarity of fake news across different languages.
- Finding trends/similarity of fake news across different countries.
- Collaboration between Fact checkers, are they working on the same claims?

Thank you!!

More details is Available at

https://gautamshahi.github.io/FakeCovid/

References

- Shahi GK, Dirkson A, Majchrzak TA. An Exploratory Study of COVID-19 Misinformation on Twitter. arXiv preprint arXiv:2005.05710. 2020 May 12.
- Shahi GK, Nandini D. FakeCovid--A Multilingual Cross-domain Fact Check News Dataset for COVID-19. arXiv preprint arXiv:2006.11343. 2020 Jun 19.
- https://www.poynter.org/ifcn-covid-19-misinformation/
- https://trends.google.com/trends/
- https://www.snopes.com/collections/
- https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes