FactDrill: A Data Repository of Fact-Checked Social Media Content to Study Fake News Incidents in India

Shivangi Singhal[1], **Rajiv Ratn Shah**[1], **Ponnurangam Kumaraguru**[2]
[1]Indraprastha Institute of Information Technology, Delhi, India
[2]International Institute of Information Technology, Hyderabad, India
(shivangis, rajivratn)@iiitd.ac.in, pk.guru@iiit.ac.in

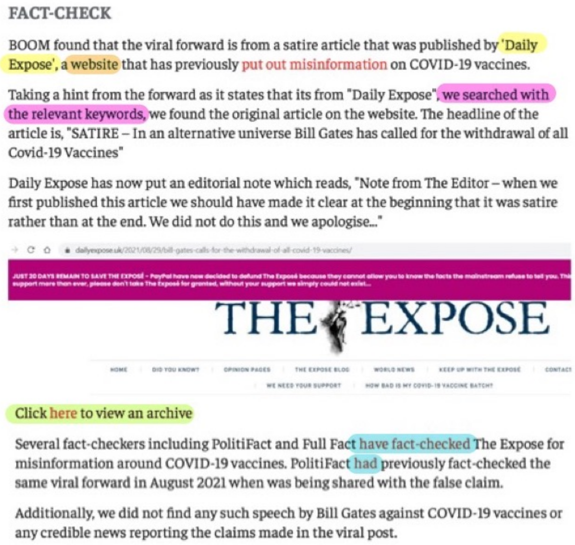
Abstract

The production and circulation of fake content in India is a rising problem. There is a dire need to investigate the false claims made in public. This paper presents a dataset containing 22,435 fact-checked social media content to study fake news incidents in India. The dataset comprises news stories from 2013 to 2020, covering 13 different languages spoken in the country. We present a detailed description of the 14 different attributes present in the dataset. We also present the detailed characterization of three M's (multi-lingual, multi-media, multi-domain) in the FactDrill dataset. Lastly, we present some potential use cases of the dataset. We expect that the dataset will be a valuable resource to understand the dynamics of fake content in a multi-lingual setting in India.

Introduction

With its massive population, the rise in production and circulation of fake content online is posing a serious social challenge in India.[1] To limit the escalation of fake news; a constant effort is made towards designing automated fact-checking (Vlachos and Riedel 2014; Ferreira and Vlachos 2016; Wang 2017; Alhindi, Petridis, and Muresan 2018; Tchechmedjiev et al. 2019; Augenstein et al. 2019) and fake news detection (Shu et al. 2018; Wang et al. 2018; Khattar et al. 2019; Zhou, Wu, and Zafarani 2020; Singhal et al. 2019; Singhal et al. 2020) solutions. However, we believe that such solutions might have a limited impact in solving the issue in India because the first language (mother tongue) of Indians is diverse and not restricted to English. As a result, we might encounter the production and distribution of fake content in the regional languages. Current datasets in English limits the ability to study the menace of fake news in the Indian context.

In this paper, we propose FactDrill: a data repository of fact-checked social media content to understand the dynamics of fake content in a multi-lingual setting in India. The dataset presented in the paper is unique due to the following reasons:



FACT-CHECK

BOOM found that the viral forward is from a satire article that was published by 'Daily Expose', a website that has previously put out misinformation on COVID-19 vaccines.

Taking a hint from the forward as it states that its from "Daily Expose", we searched with the relevant keywords, we found the original article on the website. The headline of the article is, "SATIRE – In an alternative universe Bill Gates has called for the withdrawal of all Covid-19 Vaccines"

Daily Expose has now put an editorial note which reads, "Note from The Editor – when we first published this article we should have made it clear at the beginning that it was satire rather than at the end. We did not do this and we apologise..."

dailyexpose.uk/2021/08/29/bill-gates-calls-for-the-withdrawal-of-all-covid-19-vaccines/

JUST 20 DAYS REMAIN TO SAVE THE EXPOSÉ – PayPal have now decided to defund The Exposé because they cannot allow you to know the facts the mainstream refuse to tell you. Thi support more than ever, please don't take The Exposé for granted, without your support we simply could not exist...

Click here to view an archive

Several fact-checkers including PolitiFact and Full Fact have fact-checked The Expose for misinformation around COVID-19 vaccines. PolitiFact had previously fact-checked the same viral forward in August 2021 when was being shared with the false claim.

Additionally, we did not find any such speech by Bill Gates against COVID-19 vaccines or any credible news reporting the claims made in the viral post.

Figure 1: An excerpt from our proposed FactDrill dataset depicting the *investigation_reasoning* attribute. The attribute is exclusive of the FactDrill dataset and is not present in any existing fact-checking datasets. The attribute provides minute details of the fact-checking process.

- **Multilingual information**: There are 22 official languages in India. The 2011 Census of India[2] shows that the languages by the highest number of speakers (in decreasing order) are as follows: Hindi, Bengali, Marathi, Telugu, Tamil, Gujarati, Urdu, Odia, Malayalam, and Punjabi. On the other hand, only 10.67% of the total population of India converse in English. Though the current datasets are in English, the above statistics indicate a need to shift fake news from English to other languages. Hence, the proposed dataset consists of news samples that span over 13 different languages spoken in India.
- **Investigation reasoning**: With the FactDrill dataset, we present an attribute that explains how the manual fact-checkers carry out the investigation. Figure 1 shows a

[1] https://indianexpress.com/article/india/214-rise-in-cases-relating-to-fake-news-rumours-7511534/

[2] https://en.wikipedia.org/wiki/Multilingualism_in_India

screenshot of the attribute taken from a sample of the Boom website.[3] We believe providing such information can give insights about the news story like, *(i)* social media account or website that posted the fake content (highlighted in yellow), *(ii)* platform that first encountered the fake content (highlighted in orange), *(iii)* links to the archive version of the post if the original content is deleted (highlighted in green), *(iv)* tools used by fact-checkers to investigate the claim (highlighted in pink), and *(v)* links to the supporting or refuting reports related to the claim (highlighted in blue). Such insights have the potential to drive the research towards studying the *'Nature of fake news production'* in general. The attribute is exclusive to the FactDrill dataset.

- **Multi-media and multi-platform information**: Fake news can be published in any form and on any social and mainstream platform. The curated dataset incorporates the information about media (images, text, video, audio, or social media post) used in fake news generation and the medium (Twitter, Facebook, WhatsApp, and Youtube) used for its dissemination.
- **Multi-domain information**: The previous fact-checking dataset covers information on specific topics only. For example, Emergent (Ferreira and Vlachos 2016) only captures the national, technological, and world related happening in the US whereas (Wang 2017; Alhindi, Petridis, and Muresan 2018) include health, economic, and election-related issues. In our proposed dataset, we have curated information from the existing fact-checking websites in India, giving us leverage to capture news stories of different topics and cover events that happened during the time frame.

Our contribution can be summarized as follows:

- We have curated the first large-scale multilingual Indian fact-checking data to the best of our knowledge. The dataset comprises 22,435 samples from the 11 Indian fact-checking websites certified with IFCN ratings. The samples are Hindi, English, Bangla, Marathi, Malayalam, Telugu, Tamil, Oriya, Assamese, Punjabi, Urdu, Sinhala, and Burmese. We believe that the proposed dataset can act as a prime resource to study the menace of fake news in India.
- We introduce an attribute in the feature list termed as *investigation_reasoning*. The attribute explains the intermediate steps performed by fact-checkers to conclude the veracity of the unverified claim. This is important to study because it will help us dig into the fact-checking mechanism and propose solutions to automate the process. We discuss the use cases of the curated feature and the methodology designed to extract it from the crawled unstructured data dump.
- Detailed characterization of three M's (multi-lingual, multi-media, multi-domain) in the dataset accompanied with veracity reasoning and 13 other attributes makes it a unique dataset to study.

The complete dataset is publicly available at the following link: (Dataset DOI: https://doi.org/10.5281/zenodo.5854856) (Dataset URL: https://zenodo.org/record/5854856). We also provide a datasheet for our dataset according to Datasheets for Datasets recommendations (Gebru et al. 2021) as supplementary material.

Related Work

Several datasets were released in the past that focused on automated fact-checking and fake news detection. In this section, we provide an overview of the existing fact-checking datasets. Next, we discuss the datasets that focus on India and how the proposed FactDrill dataset differs from it.

Overview of Fact-checking Datasets

Vlachos and Riedel (Vlachos and Riedel 2014) made the first effort towards this direction in 2014. The paper released a publicly available dataset consisting of facts-checked by journalist available online. The statements were picked from the fact-checking blog of Channel 4 and the Truth-O-Meter from PolitiFact. The statements mainly captured issues prevalent in US and UK public life. Apart from statements, the meta-data features like: *(i)* publish date, *(ii)* speaker, *(iii)* fine-grained label associated with the verdict and, *(iv)* URL were also collected.

In another study (Ferreira and Vlachos 2016), the data was collected from numerous sources, including rumour sites and Twitter handles. The news on the world, US national and technology were captured. For each claim, a journalist would search for the articles that are either in *support*, *against* or *observing* towards the claim. The final dataset consists of claims with corresponding summarized headings by the journalist and associated veracity label with the final verdict on the claimed statement.

Both the previously mentioned datasets were quite small in numbers. To overcome this drawback, the LIAR dataset (Wang 2017) was introduced in 2017. It consists of around 12.8K short statements curated from the Politifact website. It mainly contains samples collected from various sources, including TV interviews, speeches, tweets, and debates. The samples cover a wide range of issues ranging from the economy, health care, taxes, and elections. The samples were annotated for truthfulness, subject, context, speaker, state, party, and prior history. For truthfulness, the dataset was equally distributed into six labels: pants-fire, false, mostly false, half-true, mostly-true, and true. In 2018, Alhindi *et al.* (Alhindi, Petridis, and Muresan 2018) proposed LIAR-PLUS, an extended version of the LIAR dataset. Human justification for the claim was automatically extracted from the fact-checking article for each sample. It was believed that justification combined with extracted features and meta-data would boost the performance of classification models. Another dataset that came into existence in 2018 was FEVER (Thorne et al. 2018). It consists of *185,445* claims that were not naturally occurring but were generated by altering sentences extracted from Wikipedia.

Later in 2019, two new functionalities *i.e.* evidence

[3] https://www.boomlive.in

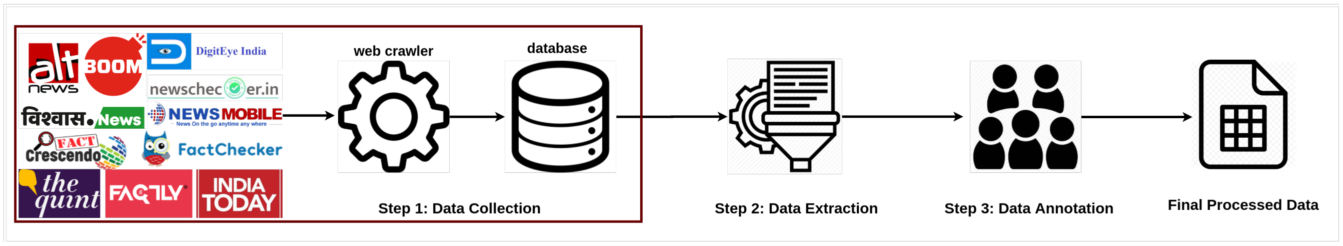


Figure 2: Our proposed dataset curation pipeline. Step 1 describes the data collection process. This is followed by Step 2, describing the data extraction methodology, and Step 3 discusses the ataata annotation and evaluation process.

pages and knowledge graph triples were introduced to fact-checking data that resulted in overall improvement of the accuracy. Augenstein *et al.* (Augenstein et al. 2019) presented the most extensive dataset on fact-checking that had 34,918 claims collected from 26 fact-checking websites in English listed by Duke Reporters Lab and on fact-checking Wikipedia page. The prime benefit introduced was the ten relevant evidence pages per claim. Other features included the claim, label, URL, reason for the label, categories, speaker, fact-checker, tags, article title, publication, date, claim date, and the full text associated with claim. On the other hand, Tchechmedjiev *et al.* (Tchechmedjiev et al. 2019) proposed ClaimsKG, a knowledge graph of fact-checked claims. It enables structured queries for features associate with the claim. It is a semi-automated method that gathers data from fact-checking websites and annotates claims and DBpedia's corresponding entities.

Overview of Fact-checking and Fake News Datasets for India

There has been little effort made to study the menace of fake news in India. Recently, Sharma *et al.* (Sharma and Garg 2021) proposed IFND: Indian Fake News Dataset, comprising of the following attributes: *(i)* title, *(ii)* date and time, *(iii)* source of the news, *(iv)* link to news, and *(v)* label. The dataset consists of 37,809 and 7,271 real and fake news samples. The real news is collected from Tribune, Times Now news, The Statesman, NDTV, DNA India, and The Indian express. The fake news samples are curated from Alt news, Boomlive, Digit eye, The logical Indian, News mobile, India Today, News meter, Factcrescendo, TeekhiMirchi, Daapan, and Afp. In another attempt, Dhawan *et al.* (Dhawan et al. 2022) proposed FakeNewsIndia to examine the fake news incidents in India. The team curated 4,803 fake news stories from June 2016- December 2019 from 6 fact-checking websites, namely, Times of India, Alt news, Afp, India Today, pIndia, and Factly. The dataset comprises the following attributes, title, author, text, video, date-time, and website. The authors have also curated 5,031 tweets and 866 Youtube videos present in the dataset.

Though the datasets have made an effort to create resources that cater Indian region, still it faces a few limitations, *(i)* The IFND dataset is highly imbalanced. No assurance about the authenticity of sources is provided, *(ii)* In the FakeNewsIndia dataset, the sample count is low. Data curation is also performed for a short period, *(iii)* Both the curated datasets consists of samples in English, missing the data in regional languages, *(iv)* There are numerous attributes present in a website but both the papers limits to some specific features. This might lead to information loss, and *(v)* None of the datasets had the *investigation_reasoning* attribute proposed in the *FactDrill* dataset.

Our proposed FactDrill dataset overcomes all the drawbacks mentioned above. Next, we discuss the data curation process.

Step 1: Data Collection

We have curated the first large-scale multilingual Indian fact-checking data to the best of our knowledge. Figure 2 shows the complete data curation process. In this section, we focus on the first step *i.e.* the data collection.

Though fact-checking services play a pivotal role in combating fake content online, little is known about whether users can rely on them or not. To corroborate trust among the audience, fact-checking services should endeavour transparency in their processes, organizations, and funding sources. To look out for trusted Indian fact-checking websites, we came across International Fact Checking Network (IFCN).[4] Next, we discuss in detail about IFCN, its measuring criteria, and sources that are chosen for preparing the proposed dataset.

International Fact-Checking Network

The International Fact-Checking Network is owned by the Poynter Institute of Medical Studies, located in St. Petersburg, Florida. It was set in motion in September 2015. The prime objective to establish IFCN was to bring together the fact-checkers present across the globe under one roof. It also intends to provide a set of guidelines through the fact-checkers code of principles mandatory for the fact-checking organizations to follow. The code of principles is designed for agencies that actively broadcast the proper investigation against the false claim on mainstream or social media platforms.

The legally registered organizations routinely scrutinize the statements made by public figures, and prominent institutions are generally granted the IFCN signatory. The statements can be text, visual, audio and other formats mainly

[4] https://ifcncodeofprinciples.poynter.org/signatories

Website	Establishment	Languages Supported	Domain
Alt News	Feb 2017	English, Hindi	Politics, Science, Religion, Society
Boom Live	Nov 2016	English, Hindi, Bangla, Burmese	General
DigitEye India	Nov 2018	English	General
FactChecker	Feb 2014	English	General, Modified
Fact Crescendo	July 2018	English, Hindi, Tamil, Telugu, Kannada, Urdu, Oriya, Assamese, Punjabi, Bengali, Marathi, Gujarati, Malayalam,	General, Coronavirus
Factly	Dec 2014	English, Telugu	General, Coronavirus
India Today		English	General
News Mobile	2014	English	General
NewsChecker		English, Hindi, Marathi, Punjabi, Gujrati, Tamil, Urdu, Bengal	General
Vishvas News		English, Hindi, Punjabi, Odia, Assamese, Gujrati, Urdu, Tamil, Telugu, Malayalam, Marathi	Coronavirus, Politics, Society, World, Viral, Health
The Quint- Webqoof		English, Hindi	General, Health

Table 1: An overview of the fact-checking sources considered during the data collection. Empty cells indicate that the information is not available.

related to public interest issues. On the other hand, organizations whose opinions look influenced by the state or any other influential identity or a party are generally not admitted to the grant.

To be eligible for an IFCN signatory, the organization is critiqued by independent assessors on 31 criteria. The assessment is then finally reviewed by the IFCN advisory board to ensure fairness and consistency across the network. There are about 82 Verified signatories of the IFCN code of principles among which 11 are based on India.[5] To ensure the authenticity and verifiability of the curated data, we have considered those Indian fact-checking sources that are IFCN rated verified.

Next, we discuss the Indian fact-checking websites considered for our data collection process.

Indian Verified Fact-Checking Websites

The prime benefit of gathering data from fact-checking websites is that we can read the reasoning behind the veracity of a news sample. The detailed description of the investigation gives valuable insight to the reader about how and why the viral claim was false. With this objective in mind, we decided to collect data from the existing fact-checking websites on a mission to debunk fake information from the Indian ecosystem. Table 1 provides an overview of the fact-checking websites considered for data curation. The table highlights the key features of a particular website in the form of, *(i)* organization establishment year, *(ii)* languages debunked by the website, and *(iii)* domain covered.

Step 2: Data Extraction

In this section, we discuss the schema of our proposed dataset. It is the second step of the data curation pipeline, as shown in Figure 2.

We set up a data extraction system that makes use of a Python library, Beautiful Soup[6] to extract data from web pages. Our system checks the sources for new data once in 24 hours. In this paper, we present a study on samples curated from 2013 to 2020. By the end of the data curation process, we had 22,435 news samples. Among them, 9,058 samples belong to English, 5,155 samples to Hindi and the remaining 8,222 samples are distributed in various regional languages *i.e.* Bangla, Marathi, Malayalam, Telugu, Tamil, Oriya, Assamese, Punjabi, Urdu, Sinhala, and Burmese.

Dataset Attributes

We curate numerous features from the unstructured data. A sample showcasing all the attributes is present in Figure 3. We have categorized the extracted feature set into various classes like meta-features, textual features, media features, social features, and event features. The final processed data is shown in Figure 4.

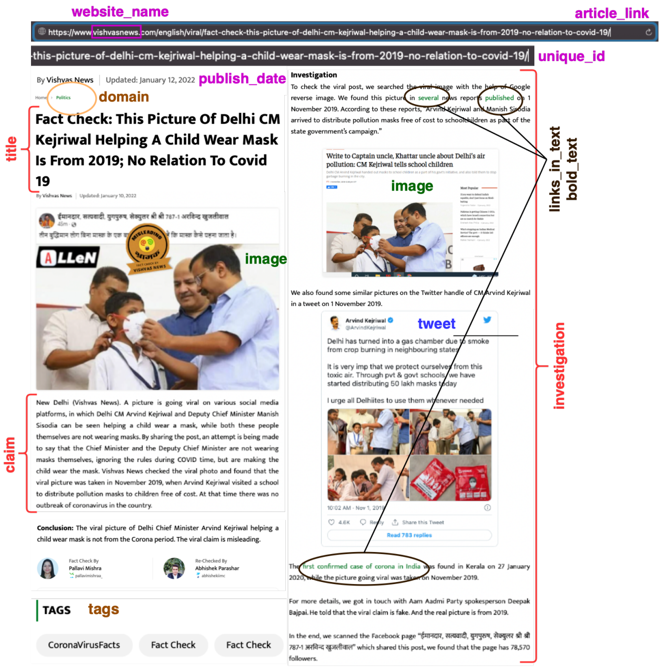
1. Meta Features

   We consider those attributes as meta_features that tells us about the sample, like *website_name, article_link, unique_id, publish_date*.

   - *website_name*: Denotes the name of the source from where the sample is collected.

[5] As per 2020 statistics.

[6] https://pypi.org/project/beautifulsoup4/



https://www.vishvasnews.com/english/viral/fact-check-this-picture-of-delhi-cm-kejriwal-helping-a-child-wear-mask-is-from-2019-no-relation-to-covid-19/

-this-picture-of-delhi-cm-kejriwal-helping-a-child-wear-mask-is-from-2019-no-relation-to-covid-19/

By Vishvas News | Updated: January 12, 2022

Home › Politics

Fact Check: This Picture Of Delhi CM Kejriwal Helping A Child Wear Mask Is From 2019; No Relation To Covid 19

By Vishvas News | Updated: January 10, 2022

ईमानदार, सत्यवादी, युगपुरुष, सेक्युलर श्री श्री 787-1 अरविन्द खुजलीवाल
45m

तीन बुद्धिमान लोग बिना मास्क के एक [illegible] कि मास्क कैसे पहना जाता है।

New Delhi (Vishvas News). A picture is going viral on various social media platforms, in which Delhi CM Arvind Kejriwal and Deputy Chief Minister Manish Sisodia can be seen helping a child wear a mask, while both these people themselves are not wearing masks. By sharing the post, an attempt is being made to say that the Chief Minister and the Deputy Chief Minister are not wearing masks themselves, ignoring the rules during COVID time, but are making the child wear the mask. Vishvas News checked the viral photo and found that the viral picture was taken in November 2019, when Arvind Kejriwal visited a school to distribute pollution masks to children free of cost. At that time there was no outbreak of coronavirus in the country.

Conclusion: The viral picture of Delhi Chief Minister Arvind Kejriwal helping a child wear mask is not from the Corona period. The viral claim is misleading.

Fact Check By
Pallavi Mishra
pallavimishraa_

Re-Checked By
Abhishek Parashar
abhishekiimc

TAGS

CoronaVirusFacts | Fact Check | Fact Check

Investigation

To check the viral post, we searched the viral image with the help of Google reverse image. We found this picture in several news reports published on 1 November 2019. According to these reports, "Arvind Kejriwal and Manish Sisodia arrived to distribute pollution masks free of cost to schoolchildren as part of the state government's campaign."

Write to Captain uncle, Khattar uncle about Delhi's air pollution: CM Kejriwal tells school children

Delhi CM Arvind Kejriwal handed out masks to school children as a part of his govt's initiative, and also told them to stop garbage burning in the city.

Most Popular

We also found some similar pictures on the Twitter handle of CM Arvind Kejriwal in a tweet on 1 November 2019.

Arvind Kejriwal
@ArvindKejriwal

Delhi has turned into a gas chamber due to smoke from crop burning in neighbouring states

It is very imp that we protect ourselves from this toxic air. Through pvt & govt schools, we have started distributing 50 lakh masks today

I urge all Delhiites to use them whenever needed

10:02 AM · Nov 1, 2019

4.6K Reply Share this Tweet

Read 783 replies

The first confirmed case of corona in India was found in Kerala on 27 January 2020, while the picture going viral was taken on November 2019.

For more details, we got in touch with Aam Aadmi Party spokesperson Deepak Bajpai. He told that the viral claim is fake. And the real picture is from 2019.

In the end, we scanned the Facebook page "ईमानदार, सत्यवादी, युगपुरुष, सेक्युलर श्री श्री 787-1 अरविन्द खुजलीवाल" which shared this post, we found that the page has 78,570 followers.

Figure 3: We present a screenshot from a fact-checking website (Vishvas News) to depict different attributes present in the proposed FactDrill dataset.

- *article_link*: The attribute gives the original link to the curated sample.
- *unique_id*: This attribute acts as the primary key for data storage.
- *publish_date*: The attribute signifies the date on which the fact-checking websites published the article.

2. Textual Features: A fact-checked article can be segregated into *title* and *content*. The *content* attribute can further be divided into *claim* and *investigation*. All the three attributes together form the textual features in our proposed dataset. Since the curated data from the website is highly unstructured, information in claims and investigation is generally present in the content part of the data. This information is extracted from the *content* attribute using human intervention. This is discussed in detail in Section Annotation Process.
   - *title*: The title of the article.
   - *content*: The attribute act as the body of the article that consist of information in the form of claim and investigation.
   - *claim*: This attribute gives the reader background information about *what was said in the related post.*
   - *investigation*: This attribute help readers in understanding *why the fact-checkers concluded a particular post to be fake*. The whole inspection process is discussed in detail with tools and techniques used to explore.
3. Media Features: The claim viral on any social media platform or mainstream media have many modalities. Similarly, the investigation to conclude the status of any viral news is also backed by numerous supporting claims that can again be in any multimedia form. The set of attributes that are categorized into multimodal features are as follows:
   - *image_links*: The links of all other images that will either belong to the original claimed images group or are presented in support of the viral claim are put under this feature as a list object.
   - *video_links*: For those samples where prime media used for fabrication is video, the link to the original video is provided by fact-checkers to back their investigation. This attribute stores all such links.
   - *audio_links*: The attribute presents all the supporting audio links related to the viral claim.
   - *links_in_text*: To provide complete justification to what was said in the investigation report, fact-checkers provide different media links in support of their investigation. All such links are present in the attribute. However, to identify where a specific link is mentioned in the fact-checked article, an attribute named as *bold_text* is used for easy identification and matching of the corresponding text from the article.
4. Social Features: The attribute stores the tweet ids present in the sample. The tweet ids can be the post that *(i)* needs to be investigated, or *(ii)* is present in support of the fake claim. We can extract the complete information from the tweet thread with this attribute.
5. Event Features: The set of features in this group gives information about the event to which a news sample belongs. These include *topic* and *tags* attributes. *For example*, the Boom article titled: 'False: Chinese Intelligence Officer Reveals Coronavirus Is A Bioweapon' had the following tags (*Coronavirus China, COVID-19, Coronavirus outbreak, Bioweapon, Biological warfare, China, Intelligence Officer*) associated with it. This kind of information helps identify the topic of the article.

Step 3: Data Annotation

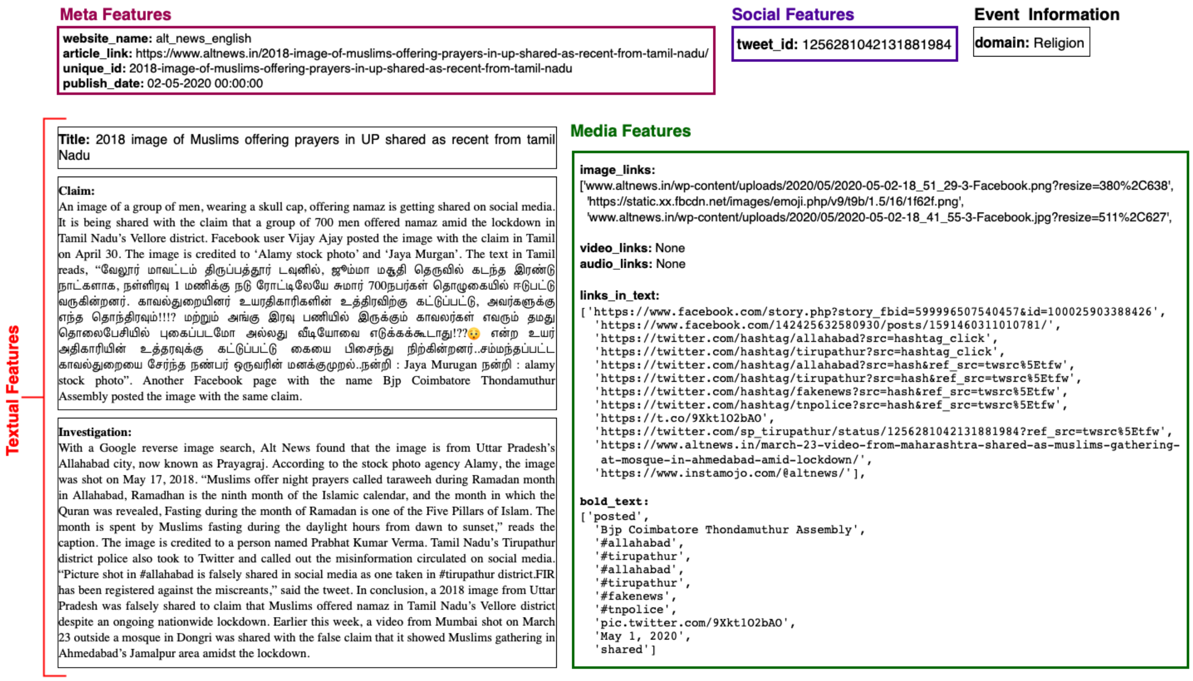
This section addresses the three key questions that facilitate the data annotation process.

Description of the Annotation Tasks

The complete annotation process is divided into two tasks. In Task 1, annotators have to mark the sample as fake or non-relevant. The non-relevant subset include samples, *(i)* that were investigated to be true, *(ii)* articles containing general fact information that news websites usually publish[7] and, *(iii)* weekly-wrap up articles that increase the chance of duplication in the dataset.

In Task 2, the annotators are provided with the three attributes, namely, content, claim and, investigation. The *content* attribute is already divided into claim and investigation

[7] https://www.boomlive.in/technologies-will-tackle-irrigation-inefficiencies-agricultures-drier-future/



Meta Features

website_name: alt_news_english
article_link: https://www.altnews.in/2018-image-of-muslims-offering-prayers-in-up-shared-as-recent-from-tamil-nadu/
unique_id: 2018-image-of-muslims-offering-prayers-in-up-shared-as-recent-from-tamil-nadu
publish_date: 02-05-2020 00:00:00

Social Features

tweet_id: 1256281042131881984

Event Information

domain: Religion

Textual Features

Title: 2018 image of Muslims offering prayers in UP shared as recent from tamil Nadu

Claim:
An image of a group of men, wearing a skull cap, offering namaz is getting shared on social media. It is being shared with the claim that a group of 700 men offered namaz amid the lockdown in Tamil Nadu's Vellore district. Facebook user Vijay Ajay posted the image with the claim in Tamil on April 30. The image is credited to 'Alamy stock photo' and 'Jaya Murgan'. The text in Tamil reads, "வேலூர் மாவட்டம் திருப்பத்தூர் டவுனில், ஜும்மா மசூதி தெருவில் கடந்த இரண்டு நாட்களாக, நள்ளிரவு 1 மணிக்கு நடு ரோட்டிலேயே சுமார் 700நபர்கள் தொழுகையில் ஈடுபட்டு வருகின்றனர். காவல்துறையினர் உயரதிகாரிகளின் உத்திரவிற்கு கட்டுப்பட்டு, அவர்களுக்கு எந்த தொந்திரவும்!!!? மற்றும் அங்கு இரவு பணியில் இருக்கும் காவலர்கள் எவரும் தமது தொலைபேசியில் புகைப்படமோ அல்லது வீடியோவை எடுக்கக்கூடாது!??😯 என்ற உயர் அதிகாரியின் உத்தரவுக்கு கட்டுப்பட்டு கையை பிசைந்து நிற்கின்றனர்..சம்மந்தப்பட்ட காவல்துறையை சேர்ந்த நண்பர் ஒருவரின் மனக்குமுறல்..நன்றி : Jaya Murugan நன்றி : alamy stock photo". Another Facebook page with the name Bjp Coimbatore Thondamuthur Assembly posted the image with the same claim.

Investigation:
With a Google reverse image search, Alt News found that the image is from Uttar Pradesh's Allahabad city, now known as Prayagraj. According to the stock photo agency Alamy, the image was shot on May 17, 2018. "Muslims offer night prayers called taraweeh during Ramadan month in Allahabad, Ramadhan is the ninth month of the Islamic calendar, and the month in which the Quran was revealed, Fasting during the month of Ramadan is one of the Five Pillars of Islam. The month is spent by Muslims fasting during the daylight hours from dawn to sunset," reads the caption. The image is credited to a person named Prabhat Kumar Verma. Tamil Nadu's Tirupathur district police also took to Twitter and called out the misinformation circulated on social media. "Picture shot in #allahabad is falsely shared in social media as one taken in #tirupathur district.FIR has been registered against the miscreants," said the tweet. In conclusion, a 2018 image from Uttar Pradesh was falsely shared to claim that Muslims offered namaz in Tamil Nadu's Vellore district despite an ongoing nationwide lockdown. Earlier this week, a video from Mumbai shot on March 23 outside a mosque in Dongri was shared with the false claim that it showed Muslims gathering in Ahmedabad's Jamalpur area amidst the lockdown.

Media Features

image_links:
['www.altnews.in/wp-content/uploads/2020/05/2020-05-02-18_51_29-3-Facebook.png?resize=380%2C638',
 'https://static.xx.fbcdn.net/images/emoji.php/v9/t9b/1.5/16/1f62f.png',
 'www.altnews.in/wp-content/uploads/2020/05/2020-05-02-18_41_55-3-Facebook.jpg?resize=511%2C627',

video_links: None
audio_links: None

links_in_text:

```
['https://www.facebook.com/story.php?story_fbid=599996507540457&id=100025903388426',
  'https://www.facebook.com/142425632580930/posts/1591460311010781/',
  'https://twitter.com/hashtag/allahabad?src=hashtag_click',
  'https://twitter.com/hashtag/tirupathur?src=hashtag_click',
  'https://twitter.com/hashtag/allahabad?src=hash&ref_src=twsrc%5Etfw',
  'https://twitter.com/hashtag/tirupathur?src=hash&ref_src=twsrc%5Etfw',
  'https://twitter.com/hashtag/fakenews?src=hash&ref_src=twsrc%5Etfw',
  'https://twitter.com/hashtag/tnpolice?src=hash&ref_src=twsrc%5Etfw',
  'https://t.co/9Xkt1O2bAO',
  'https://twitter.com/sp_tirupathur/status/1256281042131881984?ref_src=twsrc%5Etfw',
  'https://www.altnews.in/march-23-video-from-maharashtra-shared-as-muslims-gathering-
   at-mosque-in-ahmedabad-amid-lockdown/',
  'https://www.instamojo.com/@altnews/'],
```

bold_text:

```
['posted',
  'Bjp Coimbatore Thondamuthur Assembly',
  '#allahabad',
  '#tirupathur',
  '#allahabad',
  '#tirupathur',
  '#fakenews',
  '#tnpolice',
  'pic.twitter.com/9Xkt1O2bAO',
  'May 1, 2020',
  'shared']
```

Figure 4: A excerpt from the dataset displaying different attributes present in a sample of the proposed *FactDrill* dataset. The feature list is paced under different headers namely, *meta features*, *text features*, *social features*, *media features*, and *event information*. The attributes are discussed in Section Dataset Attributes.

using a keyword-based heuristic method. The role of the annotator is to check whether the segregation performed is correct or not. If not, the text is placed under the correct header.

Annotation Process

We hired language experts to annotate the samples. For Hindi and English languages, each sample is annotated by two annotators. However, due to the limited expertise in the regional languages, each sample is annotated by a single annotator. The annotators are provided with the annotation guidelines that include instructions about each task, definition of the attributes that need to derive from the text and, a few examples. The annotators studied the document and worked on a few examples to familiarize themselves with the task. They were given feedback on the sample annotations, which helped them refine their performance on the remaining subset.

Annotation Evaluation Metric

To evaluate the performance of Task 1, we calculate the inter-annotator agreements using Cohen's Kappa (Cohen 1960). We observe a mix of moderate and substantial agreement for the majority of the samples. Table 2 summarizes the Cohen's kappa measures. Though Cohen's Kappa performs exceptionally well when a dichotomous decision is involved and takes care of chance agreement, it fails badly when annotators show near 100% agreement. This phenomenon is termed as 'the paradoxes of Kappa'. During our evaluation, we observe high agreement between annotators for 1000 samples. To solve the 'the paradoxes of Kappa' issue, we used Gwet's AC(1), and AC(2) statistic (Gwet 2008). It overcomes the paradox of high agreement and low-reliability coefficients. Table 2 summarizes the Gwet's score for those 1000 samples.

To evaluate the performance for Task 2, we checked for matched ordinal positions for each annotated sample. The final inter agreement score is computed using the percent agreement for two raters (Topf 1986). It is calculated by dividing the total count of the matched sample by the total number of samples in the data.

Basic Dataset Characterization

We begin by providing a statistical overview of our proposed dataset.

Summary Statistics

Figure 5 (a) shows the distribution of samples across languages in our proposed *FactDrill* dataset. The diffusion of samples in the regional interface is majorly dominant by Bangla, Malayalam, Urdu and, Marathi language. Figure 5 (b) represents the number of samples belonging to the fact-checking websites. Among them Fact Crescendo website rules in debunking fake news dissemination in different languages.

Popular Fake Events in India

We analyze the topic distribution of fact-checking articles in different languages, *i.e.* English, Hindi and Regional lan-

Website	Language	Inter Annotator Agreement Score		# Samples
		Task 2 (Percent Agreement)	Task 1 Cohen‘s Kappa /Gwet‘s AC(1) AC(2)	
Alt News	English	0.78	0.48	2058
	Hindi	0.76	0.53	1758
Boom	English	0.42	0.66	909
	Hindi	0.90	0.53	880
DigitEye	English	0.86	0.56	147
FactChecker	English	0.31	0.15	156
Fact Crescendo	English	1.00	**1.00**	256
	Hindi	0.99	**1.00**	264
Factly	English	0.92	0.76	971
India Today	English	0.95	0.44	788
News Mobile	English	0.71	0.29	1543
Vishvas News	English	0.94	**0.91**	254
	Hindi	0.98	**0.90**	1369
Webqoof	English	0.86	0.47	1771
	Hindi	0.95	**0.97**	328

Table 2: Inter-annotator agreement for the two tasks. The values in bold indicates that Gwet's AC(1) and AC(2) scores are calculated for the samples. We observe a mix of moderate (0.41-0.60) and substantial (0.61-0.80) agreement for the majority of the samples.

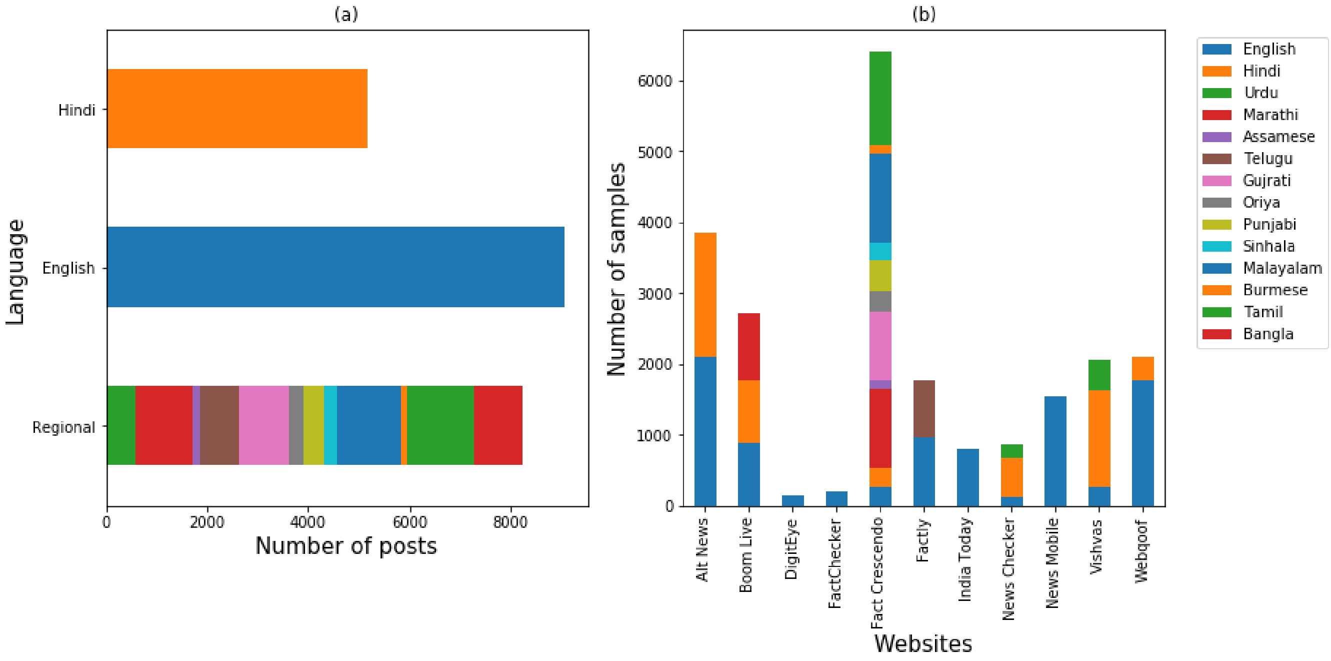


Figure 5: (a) shows the distribution of different languages in our proposed dataset, (b) shows the spread of data across different websites. It also depicts the language supported by each website.

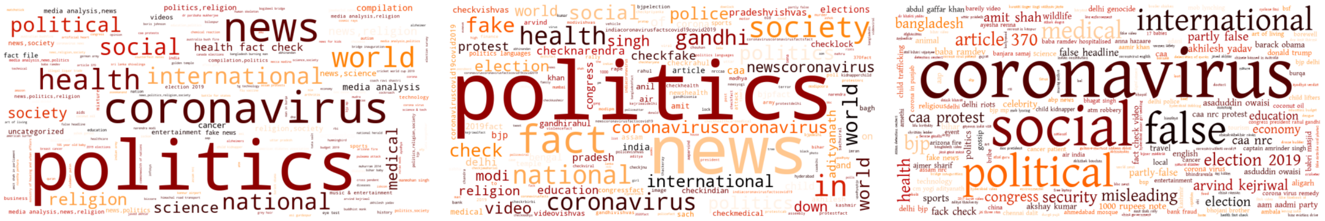


Figure 6: Topic Distribution in English, Hindi and Regional languages (left to right). The figures clearly show that most fake news dissemination across the country is centred on the political domain.

guages. All the tags and domain knowledge for the Hindi and regional languages were present in English only. From Figure 6, we can conclude that political activity is an essential ground for fake news creation in all the languages. With the onset of year the 2020, the world has witnessed a global pandemic *i.e.* Coronavirus. This has affected peoples' lives and has also given rise to the infodemic of fake content. To no surprise, the second popular topic for fake news creation in India was Coronavirus, followed by health and religion.

Circulation of Fake News in India

Figure 7 (a) shows the percentage increase in production of fake news over the years. The fact-checking trend came to India in 2013, majorly debunking news in the English Language. As and when fake news dissemination in English got little popular *i.e.* 2017, we saw it intruding in the other languages too. This steady shift to other languages was observed quite lately *i.e.* in 2017 and 2018. We observe sharp peaks and drops in the graph that will be an interesting study in the future. Figure 7 (b) shows the year-wise distribution of samples in the dataset. The graph shows a steady increase in fake news production, with a major peak observed in 2019. For both these observations, the data considered for the year *2020* is till June.[8]

Use Cases

There are varied threads of fake news research that can be initiated with the help of FactDrill dataset. We want to propose some ideas for the same formally.

- **Understanding the nature of fake news:** Various efforts have been made to-date to eliminate fake news on the Internet. There are two primary drawbacks to such approaches, 1) The system performs well on trained samples but fails drastically on real-world data, 2) the performance of classifiers varies considerably based on the evaluation archetype, and performance metric (Bozarth and Budak 2020). We believe there is a need to study the nature of fake news before attempting to detect it. Towards this end, we present a dataset that provides a detailed investigation of the fake sample that includes, *(i)* the modality faked in the news, *(ii)* 'how' the sample was concluded to be false and, *(iii)* tools used to conclude.
- **Suppressing Fake News Dissemination at an Early Stage:** Fact-checkers are making a constant effort to debunk false information online. However, we still witness duplicates and republish content online. This demonstrates that fact-checking initiatives are not reaching the general public. With FactDrill dataset, we can develop technologies stationed at different social media platforms; such systems can use information from the debunked news sample and make it available to the audiences on the platform.
- **Bias among fact-checkers:** Fact-checking is tedious. Different websites aim to debunk news of different topics. There can be websites that aim at exposing a particular kind of information. It will be interesting to look out for biases in the fact-checking pattern and its related effects.
- **Modelling Temporal Progression:** FactDrill dataset consists of data that spans from the year 2013 to the year 2020. It can serve as an excellent source to study the evolution of fake news over the years.
- **Event-centric Studies:** FactDrill dataset comprises of news stories that span different events across the timeline. For instance, it had fake news stories busted during the CAA, NRC Bill, COVID-19 pandemic, to name a few. The proposed dataset can be used to study the impact of fake news dissemination during such popular events in the country.
- **Exploring the Multilingual Fake News Direction:** FactDrill dataset comprises news stories that span over 13 different languages spoken in the country. The proposed dataset will help in designing automatic detection and language identification systems. Moreover, data in multiple languages can further open up research opportunities in the Natural Language Processing (NLP) domain.
- **Challenge Proposal:** We want to extend our work as a challenge proposal to dig deep into studying the fake news patterns in India.

FAIR Principles

Our proposed FactDrill dataset adheres to the four FAIR data principles: Findable, Accessible, Interoperable and Reusable, as follows: *(i)* The complete dataset is publicly available at the following link: (Dataset DOI: https://doi.org/10.5281/zenodo.5854856) (Dataset URL: https://zenodo.org/record/5854856) making FactDrill

[8] During the data collection stage, the last sample collected was in June 2020.

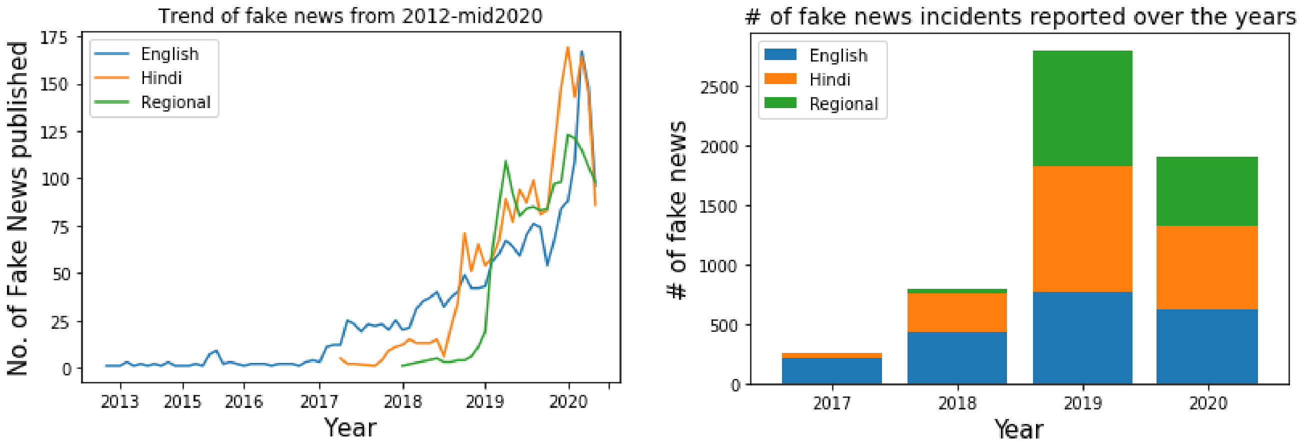


Figure 7: Circulation of fake news over the years in India. During the data collection stage, the last sample collected was in June 2020. Hence, the sample count is shown till June.

dataset easily Findable and Accessible, *(ii)* The proposed dataset is provided in an xlsx (Excel Spreadsheet) format that can be viewed and parsed easily. We have also provided a detailed explanation of each attribute in the FactDrill dataset in Section Dataset Attributes. In addition to that, it can be exported to other data formats like CSV (Comma Separated Values), making FactDrill dataset Interoperable and Reusable.

Conclusion

This paper presents FactDrill: a data repository of fact-checked social media content to study fake news incidents in India. To the best of our knowledge, this is the first large scale multilingual Indian fact-checking data that provides fact-checked stories for 13 different languages spoken in the country. We believe such a dataset can aid researchers in exploring the fake news spread in the regional languages. Additionally, researchers could also look out for the dissemination of fake content across the different language silos. The FactDrill dataset comprises 22,435 samples from the IFCN rated Indian fact-checking websites. Fourteen features associated with each sample are grouped under meta, textual, media, social, and event features. We also present a new attribute to the feature list *i.e. investigation reasoning* and explain its relevance and need in the current fact-checking mechanism. In the future, we would like to organize challenges around this data to instigate researchers in asking interesting questions, finding limitations, and proposing any improvements or novel computational techniques in detecting fake news in low resource Indian languages.

Acknowledgements

Shivangi Singhal is supported by TCS Research Scholar Program. Rajiv Ratn Shah is partly supported by the Infosys Center for AI and the Center of Design and New Media at IIIT Delhi. We also thank Rishabh Kaushal, Hitkul Jangra, Mallika Subramanian and Avinash Tulasi, members of the Precog Research Lab at IIIT-Hyderabad, for the valuable discussions.

References

Alhindi, T.; Petridis, S.; and Muresan, S. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 85–90. Brussels, Belgium: Association for Computational Linguistics.

Augenstein, I.; Lioma, C.; Wang, D.; Chaves Lima, L.; Hansen, C.; Hansen, C.; and Simonsen, J. G. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4685–4697. Hong Kong, China: Association for Computational Linguistics.

Bozarth, L., and Budak, C. 2020. Toward a better performance evaluation framework for fake news classification. *Proceedings of the International AAAI Conference on Web and Social Media* 14(1):60–71.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educatfonal and psychological measurement* 20(1):37–46.

Dhawan, A.; Bhalla, M.; Deeksha, A.; Rishabh, K.; and Kumaraguru, P. 2022. Fakenewsindia: A benchmark dataset of fake news incidents in india, collection methodology and impact assessment in social media. *Journal of Computer Communications*.

Ferreira, W., and Vlachos, A. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1163–1168. San Diego, California: Association for Computational Linguistics.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM* 64(12):86–92.

Gwet, K. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *The British journal of mathematical and statistical psychology* 61:29–48.

Khattar, D.; Goud, J. S.; Gupta, M.; and Varma, V. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, WWW '19, 2915–2921. New York, NY, USA: Association for Computing Machinery.

Sharma, D. K., and Garg, S. 2021. Ifnd: a benchmark dataset for fake news detection. *Complex & Intelligent Systems* 1–21.

Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *CoRR* abs/1809.01286:171–188.

Singhal, S.; Shah, R. R.; Chakraborty, T.; Kumaraguru, P.; and Satoh, S. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 39–47. Singapore: IEEE.

Singhal, S.; Kabra, A.; Sharma, M.; Shah, R. R.; Chakraborty, T.; and Kumaraguru, P. 2020. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence* 34(10):13915–13916.

Tchechmedjiev, A.; Fafalios, P.; Boland, K.; Gasquet, M.; Zloch, M.; Zapilko, B.; Dietze, S.; and Todorov, K. 2019. Claimskg: A knowledge graph of fact-checked claims. In Ghidini, C.; Hartig, O.; Maleshkova, M.; Svátek, V.; Cruz, I.; Hogan, A.; Song, J.; Lefrançois, M.; and Gandon, F., eds., *The Semantic Web – ISWC 2019*, 309–324. Cham: Springer International Publishing.

Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; and Mittal, A. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 809–819. New Orleans, Louisiana: Association for Computational Linguistics.

Topf, M. 1986. Three estimates of interrater reliability for nominal data. Nurs Res. 35(4):253–5. doi: 10.1097/00006199–198607000–00020. PMID: 3636827.

Vlachos, A., and Riedel, S. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 18–22. Baltimore, MD, USA: Association for Computational Linguistics.

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and Gao, J. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, 849–857. New York, NY, USA: Association for Computing Machinery.

Wang, W. Y. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 422–426. Vancouver, Canada: Association for Computational Linguistics.

Zhou, X.; Wu, J.; and Zafarani, R. 2020. Safe: Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 354–367. Springer.