The role of fact checks in counteracting disinformation on social media

CALLUM PASTUSZAK

BCompSc(Hons)

Supervisor: Marian-Andrei Rizoiu
Associate Supervisor: Simon Knight

A thesis submitted in fulfilment of
the requirements for the degree of
Bachelor of Computing Science (Honours)

School of Computer Science
Faculty of Engineering and Information Technology
The University of Technology Sydney
Australia

19 August 2022

Abstract

Recent years have seen an increase in online disinformation, and the consequences of ineffectively combating it, such as the growth of the Q-Anon movement. Fact checking is a necessary part of combating disinformation, however many aspects of the practice are not well understood, and platforms have struggled to fight the spread of disinformation. This project presents a new methodological approach to investigating the quality and reliability of information of URLs involved in fact checking on Twitter, in order to better understand the role of fact checks in counteracting misinformation on social media. This study details:

- New approach for classification based upon the content of the tweet
- Creating a method to measure the quality of URLs, specifically addressing the reliability of their information
- Identifying co-occurrences of different types of classified tweets, and analysis into their different interactions
- Analysing veracity and source alignment

Keywords: Fact checking, Misinformation, Twitter, Social media, Australian politics.

Acknowledgements

To my supervisors Marian-Andrei Rizoiu and Simon Knight, for sharing their knowledge, experience, time and effort.

To my friends and family, for helping me along the way.

Contents

List of Figures

CHAPTER 1

Introduction

1.1 Introduction

1.1.1 Motivation

Fact checks are themselves, not fully understood in that, they are an incredibly diverse interaction with countless factors influencing how they initiated and consumed. My project will investigate how fact checks have an influence in counteracting disinformation spread through social media. With a particular emphasis on user-to-user interaction on Twitter, where one individual fact checks another by sending them a URL used with the intention of debunking false information, this process is called an "organic fact check". Our intention is to investigate how effective this style of fact checking is; and to understand how often it occurs, why individuals accept/reject certain fact checks and how organic fact checks influence the spread of misinformation.

1.1.2 Objectives

Our objectives for this project are, to:

- Structure the raw tweets in our dataset, such that we can identify when a fact check occurs,
- Make a classification based upon the content of the tweet,
- Create a method to measure the quality of URLs, specifically addressing the reliability of their information,

- Identify co-occurrences of different classes of classified tweets (i.e. disinformation to debunking),
- Understand the different interactions of sources (URLs) on Twitter,
- Understanding veracity and source alignment.

1.1.3 Contribution

This project presents a new methodological approach to investigating the quality and reliability of URLs involved in fact checking on Twitter. We outline the process that we undertook, present our statistics to show the effectiveness of our results. Furthermore it provides a demonstration of the frequency and characteristics of 'organic user-driven fact checking' on Twitter.

CHAPTER 2

Literature review

Fact checking is the practice of systematically publishing assessments of the validity of claims made by public officials and institutions with an explicit attempt to identify whether a claim is factual [38]. The practise has grown in the public sphere, seeing a 900% and 2000% increase for newspapers and broadcast media respectively [38]. The increasing utilisation of social media as an outlet for political discussion and news consumption, has seen the popularisation of ‘fake news’, and consequently, fact checking’s importance in the social media sphere.

Several psychological phenomena are known to influence an individual’s openness to fact checking. ‘Motivated reasoning’ is a theory that outlines people’s tendency to dismiss information that conflicts with their pre-existing beliefs or world view, due to the psychological discomfort of shattering such perceptions [38]. Evidence has shown that political supporters of the Republican Movement in the United States are generally more inclined to this phenomenon, showing greater acceptance of information that supports their views, and dismissal of those that don’t. Additionally, conservatives in the United States have a more distrustful view of mainstream media outlets. Whereas Democrat supporters are equally receptive to information that supports and contradicts their world view.

Fact checking is a relatively uncommon occurrence in public settings. With research positing this because of social context, describes how individuals only tend to instigate fact checks when their social context encourages it [16]. A 2017 study investigated the relationship between social connections and fact checking. Observing instances of fact checks on Twitter that involved a direct reply containing a URL to the website Snopes.com and collecting the friendship data between the individuals. They observed that fact checks were effective in only 39% of cases when the two individuals did not previously know each other, and that

73% of fact checks were had mutual friends [16]. They also found evidence showing a positive correlation between the size of a fact checkers online following (compared to the user being checked) and the likeliness of the fact check being accepted [16]. However, the sole inclusion of cases using Snopes.com can be seen as a limitation. A researcher, Soroush Vosoughi performed a study on fact checking's effect towards rumour diffusion on Twitter. Their method consisted of collected a list of six fact checking URL's.

Fact checking originating from mainstream outlets has been scrutinised by research, finding a negative correlation between the 'sophisticated language' and the perceived accuracy of an article, and articles that use the so called 'implied truth effect', which involves dismissing articles as fake news. [38]. Furthermore, the use of 'truth scales', when a fact checking article used a scale to visually summarise the truthfulness of a claim, was shown not only to be ineffective, but counterproductive in convincing users of their findings[38]. Fact checking has that refutes an entire message as opposed to portions of a message, have been shown to be substantially more effective [38]. See more in [3, 4, 11, 13, 22, 23, 27, 30–32, 36, 40–42].

Major social media platforms such as Facebook and Twitter several methods to prevent disinformation spread. Facebook uses algorithms to detect possible fake news and employs certified fact checkers to review its status [6]. These fact checkers can embed warning alongside content labelled as fake news, warning users to this rating, and explaining why this was done. They also warn users who frequently post fake news stories. Twitter uses similar methods, embedding whether a link is Misleading information, Disputed or Unverified [29]. Twitters solution is relatively new, resulting in a lack of research as to its specific effectiveness, however research has shown that Facebooks warning discourage users from sharing content labelled as fake [17]. However, such a method has faults. The small teams of fact checkers cannot successfully address the sheer volume of rapidly shared, and evolving misinformation. Furthermore, platforms algorithms can severely hamper this style of fact checking, as admitted in Facebook internal documents [2]. Facebook state that they rely upon detection of doubt and conflict in discussions as a method to instigate a fact check but relent that the communities most responsible for the perpetuation of misinformation are homogeneous, thus due to motivated reasoning, unlikely to correct themselves. See more in [10, 14, 18–20, 24–26, 28, 34, 35, 39].

Misinformation is understood to spread with greater speed than truths and outpace fact checking attempts [2]. A 2018 study on Twitter showed that "It took truth about six times as long as falsehood to reach 1500 people" [37] . Misinformation related to political topics spread with more speed and reach than any other topic, followed closely by urban legends and scientific information. Another study investigating the spread of rumours on Facebook in the form of images found that individual 'organic fact checks' on viral posts were not necessarily definitive in establishing the posts nature as true or false, owing to the large volume of comments that unintentionally drown out fact checks, but contributed to the eventual establishment of factuality [5]. See more at [1, 7, 8, 14, 15, 20, 21, 33–35, 39, 43].

The spread of mis/disinformation due to bots is a contentious topic. Vosoughi's study concludes that bots as not particularly significant to the spread of misinformation, achieving the same results before, and after removing identified bots [37] . Furthermore, it is incredibly difficult to reliably identify the difference between Disinformation (the spread of misleading information with intention to mislead or further a conflicting cause) and Misinformation (the spread of information that is misleading, without a malicious intent).

Hence, I have identified several gaps in our understanding of how fact checking effects the spread of misinformation: **First, there is a lack of understanding of fact checking in an Australian context. There have been almost no studies on social media fact checking in an Australian context**, and understanding the dynamics of fact checking within Australia's unique political environment will help identify how to design more effective fact checks, and combat misinformation on Australian social media.

Second, current research does not compare the difference in effectiveness of organic user driven fact checking in online discussions compared to official fact checking accounts. Studies have been conducted on each; however, these have been isolated studies. Directly determining the effectiveness between the two could provide: organisations with a better understanding of how to engage their audience, individuals with an understanding as to their influence or new approaches, and social media platforms to adjust their methods to further combat misinformation.

Third current research does not substantially investigate if there is a difference between the effect that fact checking URLs from a wide range of sources. There is some evidence to support political supporters from separate parties reacting differently to fact checking [9]. However, this been exclusively investigated utilising American political parties and a small number of mainstream sources. There has been little research comparing the impact of a wide range of fact checking sources, including the non-mainstream.

Fourth, can the effectiveness of organic fact checks be correlated with an effect on the speed and spread of misinformation? Current research focuses on determining the spread of rumours and analysing how a fact checking URLs caused any effect, or determining if a fact check elicited a individual positive or negative reaction. However, we do not have a great understanding of how a fact check is effective on an individual level.

2.1 Thesis Statement:

My thesis is that organic fact checking influences the spread of misinformation online, however, is dependent on a number of observable factors such as, the social context that it is used, and the source utilized to refute a claim. This study will be done on data pertaining to Australian social media. To achieve this, we address a number of research questions:

2.1.1 Question 1

Is there a way to extract and classify fact checks from an existing dataset of tweets, based upon their usage of a URL?

2.1.2 Question 2

Can we create a method to measure the quality of URLs, specifically addressing the reliability of their information?

2.1.3 Question 3

How can we determine if organic fact checking can have an effect on the overall speed and spread of misinformation?

CHAPTER 3

Methodology

3.1 Data

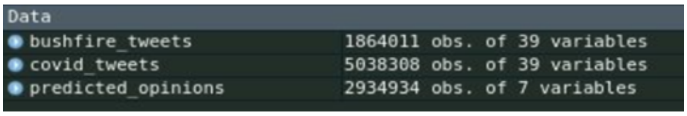
3.1.1 Initial Dataset Description

The Twitter dataset was collected by my supervisor Marian-Andrei Rizoiu and one his students Quyu Kong, for a paper classifying the opinions of online discussions [12]. It was collected from Twitter, Facebook, and YouTube, using a series of keywords of discussions about the Australian Bushfires 2019-20 and COVID-19 Pandemic in 2020. These 2 topics formed separate datasets that were identical in structure. The former dataset was collected from December 2019 – February 2020, and the latter was collected from March - May 2020.

We limited our usage of the data to only the entries sourced from Twitter, as it was determined to be too complicated as each platform facilitated different types of user interaction, and it would have required an extensive amount of work to facilitate a meaningful comparison across the platforms. Thus we ignored the Facebook and YouTube elements. Through analysis of the dataset, [12] produced an additional set of labels classifying the opinions of each tweet present in the dataset. This was treated as a supplementary dataset, that we utilized throughout the pre-processing stages. Note: Our data was collected on Twitters v1 API.

Dataset	Number of Tweets	Unique users
Bushfire:	1,864,011	119,490
COVID-19	5,038,308	254,865

Dataset	Classified Tweets	Unique opinions
Opinion:	2,934,934	55



```
Data
bushfire_tweets          1864011 obs. of 39 variables
covid_tweets             5038308 obs. of 39 variables
predicted_opinions       2934934 obs. of 7 variables
```

FIGURE 3.1: The size of our dataset

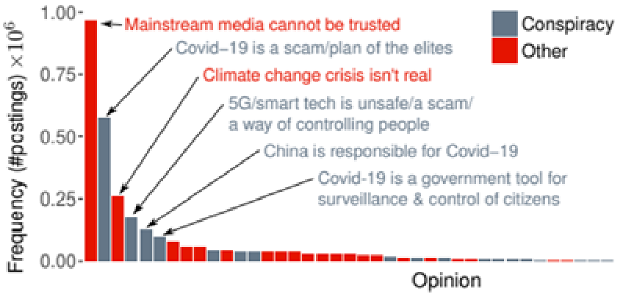


FIGURE 3.2: The distribution of opinions [12]

3.1.2 Initial Pre-processing

Both the COVID-19 and Bushfire datasets come in the form of a RDS file containing the entire JSON data for each individual tweet, and the Opinions come in a CSV. The 3 datasets were linked together, to filter complete Tweets by opinion. Additionally, this is where the Facebook and YouTube elements were filtered out.

The Opinion dataset contains duplicates of some tweets, as the classifier can assign multiple opinions to the same tweet. To deal with this I filtered tweets by their unique ID.

3.2 Problem 1: Identification of Tweets involved in Fact Checking

3.2.1 Aims and Objectives Problem 1:

- To structure the raw tweets in our dataset, such that we can identify when a fact check occurs
- To somehow make a classification based upon the content of the tweet

3.2.1.1 Evaluation Criteria:

Our dataset was not collected specifically for this project, and as our definition for fact checking requires the inclusion of a URL, this could lead to fewer fact checks being present for analysis. However this method does have some positives, as we are seeking more to understand the practise of fact checking, along with its frequency and effectiveness. This is due to our dataset containing tweets that have been classified as part of broad range of topics and opinions that are controversial, which should ensure some proportion of both dis/misinformation tweets, and fact checking tweets; as opposed to prior research, which constructed their datasets by specifically collecting tweets that would include specific URLs from a fact checking organisation [37]. We believed this method would not allow us to gauge a measure of the frequency of fact checking in controversial discussions, and close the door for investigating quasi-fact checks, such as a URL that is considered highly reliable, but may not be part of international fact checking organisation lists. It may be useful to understand these types of fact checks, as we have seen that these official fact checkers simply do not have the numbers, reach, time or ability to meet all disinformation on twitter. Hence, this project would need to extract as much information from our dataset, as possible, in order to fulfill our wishes for an investigation into more organically occurring fact checks. This is why we consider information completeness as our primary metric for success. To achieve this, we would need to design a creative, yet effective solution to identify the process of fact checking, which would open the door for investigation into our remaining research questions.

3.2.2 Methodology Problem: Classifying URLs

Our initial sourcelist for identifying a fact check was based upon the work of [37], where they created a list of certified fact checking organisations, and identified their presence in Tweet cascades. These websites were "snopes.com", "firstdraftnews.org", "politifact.com", "factcheck.org", "truthorfiction.com", "hoax-slayer.com" and "urbanlegends.about.com". We decided to append 2 further organisations onto this; abc.net.au/news/factcheck" and "poynter.org/media-news/fact-checking", thus creating our initial URL sourcelist.

3.2.2.1 Pre-Processing:

Due to the diversity of URLs, and many instances of domain shorteners, we were required to pre-process the URLs before any analysis could be performed. We utilised the R library "urltools" to further process the URLs down into domains, so that they could be more easily investigated. Our method for processing the URLs was as follows:

1. Grab the entire URL entity which is stored as an R list.
2. Unwind the `extended_url` sub-entity. This entity included the full version of any URL, mitigating many automatic shorteners.
3. Filter out "twitter.com/i/web/status", "twitter.com", "mobile.twitter.com". As we are only wishing to investigate external URLs, and wish to reduce confusion with irrelevant internal links.
4. The 'domain' function from urltools was applied to the `unwound_url`, providing us with the domain name of the website called.
5. Remove the "www." so it would conform with our sourcelist.
6. The tweets were run against our sourcelist and any matches detected are pulled out for investigation.

This reduced the size of each of our datasets from:
Bushfire: 1,864,011 to 105,611 individual tweets.
COVID-19: 5,038,308 to 261,964 individual tweets.

3.2.2.2 Initial Data Exploration

We performed an initial exploration into the data looking for instances of URLs from our fact checking sourcelist taken from [37], and modified further. This was conducted on the COVID-19 dataset, which is the larger of our 2 sets. Furthermore we limited our investigation to a subset of tweets containing the 4 most popular labeled opinions ["Mainstream media cannot be trusted", "Covid-19 is a scam/plan of the elites", "Climate change crisis isn't real", "5G/smart tech is unsafe/a scam/a way of controlling people"], due to our belief that discussion on more controversial topics and emerging topics would present more opportunities for fact checking to occur.

Prior to URL processing this subset numbered 77,762 tweets. After URL processing this number decreased to only 4,354 tweets with URLs to external websites. Of these, only 6 tweets contained instances of fact checking URLs from our sourcelist. Below is the text and accompanying URLs for each of the 6 Tweets, with any user mentions censored:

1. TEXT: https://t.co/kRHAlqvT7m has always has been a great source for fact checking most of the crap you find on the inter... anonymised
Expanded URL: https://www.snopes.com/

2. TEXT: Was Charles Lieber Arrested for Connections to Coronavirus, Wuhan Lab? #coronavirus yeah yeah nothing to see here! https://t.co/FqAhlIwsDs
Expanded URL: https://www.snopes.com/fact-check/charles-lieber-arrested-coro

3. TEXT: @– @– @– Context is important there - https://t.co/zPP5jo07Br He likened it to their... anonymised
Expanded URL: https://www.snopes.com/fact-check/trump-coronavirus-rally-rema

4. TEXT: Coronavirus Stimulus Payment Phishing Scam Email - Hoax-Slayer https://t.co/cdJwF1cVfq #phishing #cornoravirus anonymised

Expanded URL: https://www.hoax-slayer.net/coronavirus-stimulus-payment-phi

5. TEXT: @– @– Why do you believe everything on the internet? https://t.co/bIlfv9UrZm

Expanded URL: https://factcheck.afp.com/police-dismiss-false-claim-austral

6. TEXT: 'STANFORD HOSPITAL ADVICE' IS FAKE... Viral Social Media Posts Offer False Coronavirus Tips https://t.co/V1CgDA4gOJ via @factcheckdotorg

Expanded URL: https://www.factcheck.org/2020/03/viral-social-media-posts-c
?utm_source=twitter&utm_medium=social&utm_campaign=social-pug

Only 0.1378% of URLs contained links to our fact checking sourcelist, a disappointingly low number of detection's. We believed that part of the reason for this was that our initial sourcelist was comprised of mostly international fact checking organisations, and may be less popular in Australia. Further analysis of the URLs comprising this subset revealed a number of interesting features, below are the 10 most frequently linked domains:

youtube.com	624
theguardian.com	245
N/A	120
change.org	117
abc.net.au	114
nytimes.com	91
worldometers.info	70
smh.com.au	57
news.com.au	55
ozfeed.com.au	39

This allowed us to identify 3 substantial problems: 1.During the URL pre-processing, the domain shortners were invalidating a large number of URLs, meaning we would need to improve our method. 2.The domain shortening was not allowing us to view sub-domains, which is important for websites such as "abc.net.au", which has the primarily "/news" subdomain and their "news/factcheck" subdomain which is of high interest to our project Meaning we would need to further improve our URL handling. 3. Almost half of the most frequently detected domains were different sources of news media, the above domains comprise 13.045% of the subset of tweets with URLs. This opened the opportunity for further analysis of URL quality.

3.2.2.3 Methods

The poor rate of detection meant we would require an improved sourcelist. Such a sourcelist would be required to feature more diverse URLs from numerous media organisations and popular online domains, and provide a justified classification upon their adherence to fact checking standards.

Surprisingly public lists of providing such classifications are somewhat limited. Through some investigation, we found the Wikipedia perennial source list (see appendix). This list contains recommendations for Wikipedia editors about the reliability of various online sources, and their adherence to fact checking practices. Domains are given a recommendation, and a justification. They are then assigned one of 5 classes [Generally Reliable, No Consensus, Generally Unreliable, Blacklisted, or Depreciated]

The perennial list contains 354 recommendations on mainstream news outlets, known sources of disinformation, popular culture websites, satire websites and forum type websites. This amount of verity was particularly advantageous for our purposes, as we hypothesized that users may use a variety of different sources to back up their fact checks, and these may not be exclusively news sources.

We decided to supplemented Wikipedia's list with a collection of 134 international fact checking domains, certified by Poynter's International Fact-Checking Network (IFCN). We decided to create a new class for these domains called [Debunking], which we defined as only containing "Certified fact checking organisations". Our reasoning behind modifying the Wikipedia list was that, the Perennial list included both fact checking and non-fact checking sources in the same category of [Generally Reliable], yet we wished to make observations specifically for fact checking domains, while not wishing to discount highly reliable, yet not certified organisation.

Addressing our initial sourcelist, the domains "afp.com" (Agence France-Presse), "politifact.com" and "snopes.com" appeared on both the Wikipedia (as [Generally reliable]), and IFCN sourcelist; we removed the duplication and labeled them [Debunking]. The domains "truthorfiction.com" and "hoax-slayer.com" did not appear in either the Perennial sourcelist or fact checking sourcelist, but we labeled them as [Debunking] due to their utilisation in prior research investigating fact checking [37].

Further modifications: Classified RMIT ABC Fact Check abc.net.au as [Generally_reliable] due to issues capturing sub-domains.

This finished our sourcelist, with a total of 489 domains classified under 6 domain labels.

- `Debunking` Certified fact checking organisations.
- `Generally Reliable` in its areas of expertise: Editors show consensus that the source is reliable in most cases on subject matters in its areas of expertise. The source has a reputation for fact-checking, accuracy, and error-correction, often in the form of a strong editorial team.
- `No Consensus`, unclear, or additional considerations apply: The source is marginally reliable (i.e. neither generally reliable nor generally unreliable), and may be usable depending on context. It may be necessary to evaluate each use of the source on a case-by-case basis while accounting for specific factors unique to the source in question.
- `Generally Unreliable` Editors show consensus that the source is questionable in most cases. The source may lack an editorial team, have a poor reputation

for fact-checking, fail to correct errors, be self-published, or present user-generated content.

- `Blacklisted` Due to persistent abuse, usually in the form of external link spamming, the source is on the spam blacklist or the Wikimedia global spam blacklist.
- `Depreciated` Deprecated: There is community consensus from a request for comment to deprecate the source. The source is considered generally unreliable, and use of the source is generally prohibited.

Debunking	=	138
Generally Reliable	=	119
No Consensus	=	78
Generally Unreliable	=	97
Blacklisted	=	18
Depreciated	=	29

For additional testing purposes, we introduced a simplified class system, binning our existing classifications into 3 separate classes, 1.[Debunking] which contained exclusively fact checking domains, 2.[Disinfo] (Disinformation) which contained domains from the Blacklisted and Depreciated classes representing extremely low reliability websites. 3.[Other] which contained domains from 'Generally Reliable', 'No Consensus' and 'Generally Unreliable' and represented domains who's reliability is highly contextual, dependent on the contents quality.

This necessitated modification to the final part of our URL pre-processing. The final classification stage now performed as such:
6. The tweets were run against each sourcelist [Debunking], [Disinfo], [Other] and are assigned a class respective to the sourcelist they are detected in. If they are not detected in the sourcelists at all, they are assigned to the 'other' class, indicating their absence from our sourcelist of reliable websites.

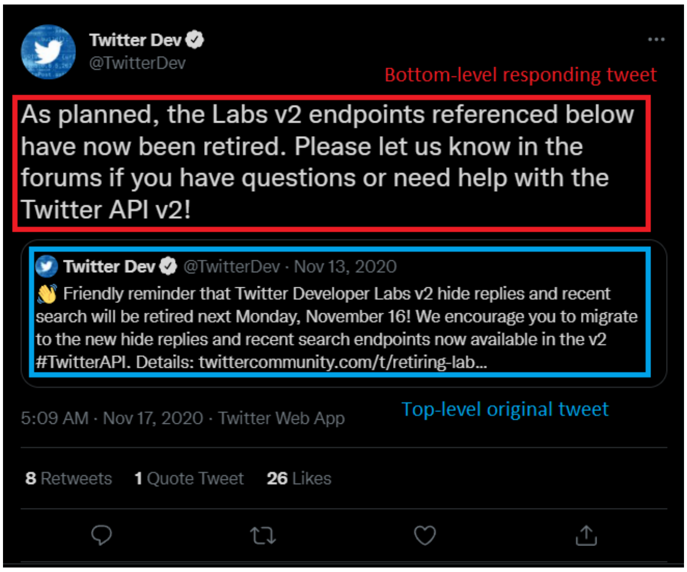
3.2.3 Methodology Problem: Investigating tweet types

3.2.3.1 Method

We had previously determined that the inclusion of a URL within a Tweet would be a good indicator as to whether it was a fact check, because a requirement of a fact check is the inclusion of an external source. Thus we would need to filter the dataset by Tweets that contained URLs. However this method involved some complexity due to the way in which Twitter stores entities for the 4 different Tweet types. When a Twitter user wishes to have an interaction on the platform, they can create their own unique Tweet (1), they can augment these posts with Hashtags(#) which Twitter considers as categories. Users can also Reply (2) to a tweet. The reply will appear in respect to the original tweet, under separate reply section. Replies can also be nested, as in; an original tweet can have replies A,B&C for which B can have its own replies X&Y, replies X&Y will appear in respect to a nested version of the original tweet and reply B. Users can also perform an action called a Retweet (3), which "reposts" the original tweet and shows the information of the user who chose to retweet it, but this does not allow for any more interaction. Another option is called a Quote Tweet (4) which does the same as a retweet, but allows the user to add some text.

Twitter stores all tweet types in effectively the same JSON structure. Each tweet contains the content of the tweet and data about the user including; unique IDs, their screen-name, the date, how many times it was retweeted, quoted, liked, favorited, etc. It also contains the object called "entities" which includes and hashtags, user mentions, etc, but crucially for us, an array storing any information relating to the presence of URLs. Furthermore, some tweet types contain entire copies of tweets they are responding to, which in turn have entities.

To ease the understanding of our intentions for pairing a tweet and its response, we created a naming scheme. As each tweet X collected within the data (regardless of type), is nested in such a way that tweet X appears first, and the tweet it is responding to is contained within, we came up with the descriptor **"bottom-level"** for tweet X, and **"top-level"** referring to any tweet contained within tweet X. This makes sense when thinking about how a tweet contained within any given tweet X, is the original post, hence when viewing on the website, the original post is always at the top, above the response.



Twitter Dev
@TwitterDev

As planned, the Labs v2 endpoints referenced below have now been retired. Please let us know in the forums if you have questions or need help with the Twitter API v2!

Twitter Dev @TwitterDev · Nov 13, 2020
👋 Friendly reminder that Twitter Developer Labs v2 hide replies and recent search will be retired next Monday, November 16! We encourage you to migrate to the new hide replies and recent search endpoints now available in the v2 #TwitterAPI. Details: twittercommunity.com/t/retiring-lab...

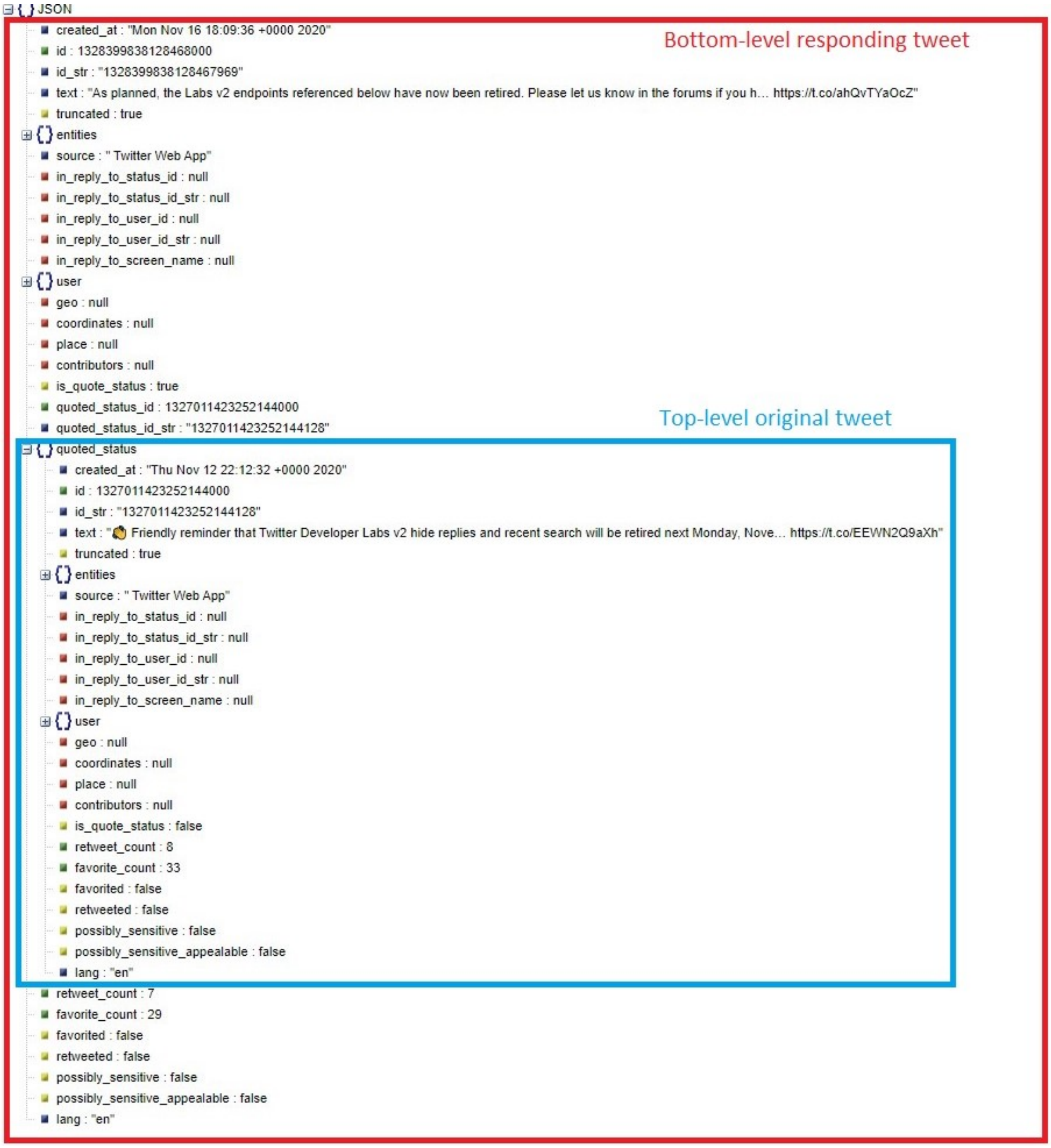
5:09 AM · Nov 17, 2020 · Twitter Web App

8 Retweets 1 Quote Tweet 26 Likes

FIGURE 3.3: Visualisation of naming scheme in the user interface

Twitter includes markers to differentiate the Type any given tweet: If the tweet is a Reply, the field `in_reply_to_status_id_str` will contain the `id_str` of the tweet it is replying to. If the tweet is a Retweet, the field `retweeted_status` will contain a copy of the top level tweet. The text will also start with "RT @" and the twitter username of the original post will be after the "@". If the tweet is a Quote, the field `is_quote_status` is set to `true`, and `quoted_status` will contain a copy of the top level tweet. Additionally, if a user is to Retweet a Quote, the tweet will follow the same structure as if it were a Retweet, but inside `retweeted_status` the subsequent top level tweet will follow the same structure as if it were a Quote. Twitter describes this case as a "Quoted Retweet".

Hence we were required to build different methods to pull the tweet entities for each tweet type, and for their corresponding top level tweets. This led to us grouping tweets by the type of tweet, [Retweets, Replies and Quotes (including Quoted Retweets)]. This is because we could pair a tweet along with what it was replying/in reply to; and allow us to observe the



```
{ } JSON
    created_at : "Mon Nov 16 18:09:36 +0000 2020"
    id : 1328399838128468000
    id_str : "1328399838128467969"
    text : "As planned, the Labs v2 endpoints referenced below have now been retired. Please let us know in the forums if you h... https://t.co/ahQvTYaOcZ"
    truncated : true
  { } entities
    source : " Twitter Web App"
    in_reply_to_status_id : null
    in_reply_to_status_id_str : null
    in_reply_to_user_id : null
    in_reply_to_user_id_str : null
    in_reply_to_screen_name : null
  { } user
    geo : null
    coordinates : null
    place : null
    contributors : null
    is_quote_status : true
    quoted_status_id : 1327011423252144000
    quoted_status_id_str : "1327011423252144128"
  { } quoted_status
        created_at : "Thu Nov 12 22:12:32 +0000 2020"
        id : 1327011423252144000
        id_str : "1327011423252144128"
        text : "👋 Friendly reminder that Twitter Developer Labs v2 hide replies and recent search will be retired next Monday, Nove... https://t.co/EEWN2Q9aXh"
        truncated : true
      { } entities
        source : " Twitter Web App"
        in_reply_to_status_id : null
        in_reply_to_status_id_str : null
        in_reply_to_user_id : null
        in_reply_to_user_id_str : null
        in_reply_to_screen_name : null
      { } user
        geo : null
        coordinates : null
        place : null
        contributors : null
        is_quote_status : false
        retweet_count : 8
        favorite_count : 33
        favorited : false
        retweeted : false
        possibly_sensitive : false
        possibly_sensitive_appealable : false
        lang : "en"
    retweet_count : 7
    favorite_count : 29
    favorited : false
    retweeted : false
    possibly_sensitive : false
    possibly_sensitive_appealable : false
    lang : "en"
```

FIGURE 3.4: Visualisation of naming scheme in the tweet object

interaction based on the URL type. We also determined 2 further conditions:

1. As fact checking necessitates a reply with some form of text, we are primarily interested in 'Replies', 'Quotes' and 'Quoted Retweets', as there is a guaranteed text interaction between 2 users. The types 'Tweet' or 'Retweet' will be temporarily ignored.

2. We determined that the fact check would require at least 1 URL. Instances where both tweets contain fact checks are preferable.

The resulting list allowed us to observe pairs of tweets that contained URLs, thus completing our first aim our allowing us to structure tweets in such a way that we can observe a fact check.

We encountered a quirk of Twitters v1 API, in that unlike all the other tweet types that have a text interaction between a post and a response, Replies do not contain a full copy of its respective top level tweet, meaning we had to figure out a way to find the matching original post. The Reply field `in_reply_to_status_id_str` acts as a pointer to the `id_str` of the tweet it is replying too, meaning we had to separately filter all instances of tweets that were of type Reply, extract their pointer, and search through the dataset to find a a corresponding `id_str`. During the data collection process, the tweets to which Replies were responding to, were not intentionally collected, meaning the Replies in our dataset were not guaranteed to have a corresponding top-level tweet, and the Reply, Quote and Retweet count values were not accurately store. After accounting for this in the filtering, out of the initial 22,278 replies, we only had 4,586 matching pairs, further filtered down to 784 matching pairs with URLs. This was an unfortunate loss of data. Furthermore, these processes were very computationally intensive, and the process of discovering the precise way to handle all the different tweet types slowed our progress greatly. To ease the strain on our computers, we trimmed the JSON files for our tweets such that only necessary data relating to the `URL`, `usernames`, `text`, `id_str` and same data respective to the responding tweet were retained.

3.2.4 Results: Fact Check Identification

COVID Total	Number
Classified Bottom:	5,967
Classified Top:	255,997

Type	Number	Set
Quotes:	5829	top and bottom
Replies:	138	top and bottom
Retweets	230,702	top

BUSHFIRE Total	Number
Classified Bottom:	2,327
Classified Top	103,284

Type	Number	Set
Quotes:	2296	top and bottom
Replies:	31	top and bottom
Retweets	91,360	top

Classified Top contains: Quotes, Replies, Retweets
Classified Bottom contains: Replies, Quotes

Our method for fact check identification massively reduced the overall size of our dataset. This was somewhat expected, as the dataset was not primarily collected with the intention of this analysis. However the number of replies was much lower than expected. This was a consequence of the data collection process. We next wanted to investigate the differences in virility of disinformation and debunking tweets. Due to the limited amount of classification with matching tweets, this analysis was only conducted on the COVID dataset, due to it being larger and having more concrete classifications

The previously discussed investigation into the instances of top level Disinfo/Debunking to a classified response tweet hasn't shown favourable results:
41 total instances within the dataset. But of these, only a total of 12 were usable. 28 were from the Reply dataset, and due to the nature of having to collect them, only one of these did not have N/A values for its reply/quote counts.
13 usable results. one instance was debunking to other, ten were disinformation to other,

and two were disinformation to disinformation Furthermore, none of these had any response counts available due to the aforementioned quirks of the reply data type.

Class Top	Class Bottom	Number of Replies:
other	disinfo	15
other	debunk	9

Class Top	Class Bottom	Number of Quotes:
other	disinfo	5
other	debunk	3

We also computed the Empirical cumulative distribution for any instance of a tweet with a URL. This distribution will show us the probability that a tweet from types Reply, Quote or Retweet; belonging to class Debunk, Disinfo, Other; having x number of interactions. These interactions consist of pulling the quotes, replies and retweets counts for the classified tweets from both datasets, and plotting the results.

(See Figures for ECDF Plots)

The graphs generally depict debunking tweets as having a lower probability of gaining a large reaction/interaction from other twitter users. This was even more noticeable in regards to top level tweets, which consistently saw a much higher rate of retweets, replies, and quotes for URLs classified as disinformation, even eclipsing that of the class ’other’ in the case of COVID top level quote counts. This finding is consistent with other studies that have shown Disinformation tweets to have a greater tendency towards becoming viral, than debunking/fact checking tweets [37]. It appeared that debunking URLs in bottom level posts did gain more reactions.

The preceding analysis has shown that this method can identify fact checks, but is reliant on the extensiveness of the URL sourcelist. Furthermore, we have seen that the conclusions of prior research regrading how disinformation is more viral than facts, hold true in our dataset. We were also able to observe the possible indications of fact checking behaviour gaining a significant response rate in bottom URLs, but these instances were quite limited.

This avenue of investigation left us at something of a dead-end due to the low amount of available classification and difficulty analysing co-occurrences on mass. We believed that we would need to further extract some meaningful measure with the existing URLs. This led us to suppose: Is there a more robust way in which the quality of the URL can be determined, and a method in which source alignment could be calculated?

3.3 Problem 2: Information quality

Our assumption was that, a twitter user performing their own fact check is not required to adhere to a strict set of fact checking criteria, and hence may be more flexible in their choice of URL for the fact check. Naturally some users may cite more reliable sources, and others, blatantly unreliable sources. Furthermore, if there was some way of detecting URLs that are identified as frequent spreaders of mis/disinformation, we could investigate how often they are fact checked. Hence, we decided to explore the idea of a class based URL classification system. As in: classifying the reliability of URLs and their adherence to fact checking standards, to better improve our fact checking detection.

3.3.1 Aims and Objectives Problem 2:

Our analysis had unearthed that the sourcelist we currently used was too restrictive to meet our criteria shift from 'the inclusion of official fact checking websites' to "the classification to URL reliability". Hence, we would need to:

- Create a method to measure the quality of URLs, specifically addressing the reliability of their information
  - This should be inclusive of the same types of URLs as our prior analysis, but also cover social media links.
- To identify co-occurrences of different classes of classified tweets (i.e. disinformation/debunking)
- Understanding the different interactions of sources(URLs) on Twitter
- Understanding veracity and source alignment

3.3.2 Methodology Problem: Class Creation

After the prior analysis, we determined that using the combination of Wikipedia's Perennial Source classifications and our additional class, were not producing satisfactory classifications. We believe this is due to the slightly different aims of the two source-lists and we decided to optimise our sourcelist to allow us to more efficiently examine the differences in-between Disinformation and Debunking. Our primary issue was, although we had identified and filtered the data into 3 distinct classes:

- 1 Debunking (fact checking organisations)
- 2 Disinformation (Known organisations that consistently spread dis/mis-information)
- 3 Disputable reliability (Domains who's content was either inconsistent in its reliability, or not relevant in the context of fact checking)
- 3.1 This also included credible information from reliable sources, that were not fact checks

The vast majority (e.g. our COVID dataset, 95.1% for Top and 94.4% for Bottom) of our data was in the third class. This greatly diminished our ability to analyse features of fact checking in more detail. Furthermore, our prior methods of extracting URLs appearing in our sourcelists, and visually analysing their alignment was inefficient. Hence, we decided that a measure of URL alignment by calculating the source veracity and measuring their agreement was a more suitable solution.

We sourced a paper from one of our research colleagues that detailed the process of measuring the influence of social media users, along with checking the veracity/reliability scores of URLS they post. This paper used a relatively similar method to ours, in that they used a URL sourcelist, to rate the URLs appearing in tweets. They used 2 sourcelists to achieve this, firstly the CoAID dataset which contains a list of specific URLs to social media posts, spreading disinformation, and secondly, a High Quality Health Sources (HQHL) that contain links to National Public Health Institutions, Prominent Health Institutions, University Websites and Medical Journals.

They also had a similar target dataset; one focusing on COVID-19 related tweets collected during August 2020, and another containing tweets relating to discussion around the Australian bushfires collected from November 2019-January 2020. The similarities in method and dataset encouraged us to utilise the methods that my colleague demonstrated in order to fix our information quality classification problem.

Their method for calculating veracity scores for domains was very similar to our method involving the perennial sourcelist, in that they checked if a specific URL is present within the CoAID dataset, and assign it the respective classification; if it is not present, then the domain is checked against the HQHS list. If neither of those steps result in a classification, then the URL is assigned 'N/A'. The veracity scores are between (1, -1) with measure of 1 indicating a ground truth of "fully reliable" URL, and -1 indicating a domain that is "fully unreliable and spreads misinformation". It also included domain level identification for a number of consistently reliable sources; which were put into the HQHL sourcelist, and were assigned a Veracity score of 1.

3.3.3 Pre-Processing

We supplemented the HQHL with our list of Fact checking domains, and specified further 4 classes to represent the different levels of reliability in domains that would not be depicted in their URL dataset. The veracity scores were assigned with guidance from my supervisors:

Classification	Veracity Class	Veracity Score
Debunking	HQSH	1
Generally Reliable	GRSL	0.75
No Consensus	NCSL	0.25
Generally Unreliable	URSL	-0.25
Blacklisted, Depreciated	DSL	-1

(Note: the veracity class names are abbreviations of the classification, followed by an abbreviation of 'sourcelist' as 'SL')

Due to the papers heavy focus on COVID-19 related information, we decided to perform our testing only utilizing our COVID dataset. However it required substantial pre-processing to conform to the codes calculations. We were required to obtain the "user id str" (renamed to author), which is a much more stable data type than "screen name" (what we were utilising at the time). This itself required us to re-run the entirety of the pre-processing outlined in problem 1, which was time consuming. We found that unwound url was the only way to process past the numerous URL shorteners.

Our classified sets thus contained:

```
'id_str', 'text', 'quote_count', 'reply_count',
'retweet_count', 'class', 'type', 'author_id', 'author_name',
'unwound_url', 'point_to_id', 'expanded_url'
```

3.3.4 Methodology Problem: Validation

We also wished to validate the efficacy of using the perennial source list to measure reliability. This would require a comparison with an external sourcelist dealing with URL reliability, and ensuring that the classifications of domains made by said external sourcelist aligns to our hybrid sourcelist.

Through a comparison, there were only minor differences noted between the Wikipedia perennial sourcelist+Fact checking list, and the CoAID dataset. These differences were also of a fairly insignificant nature, as they only dealt with the classes of Generally reliable, No Consensus and Generally unreliable.

Overall we believe this substantially validated belief that the Wikipedia Perennial Sourcelist made classifications that were reasonable, and applicable to discerning the quality of a URL.

- CoAID 205 Domains
- Wikipedia 354 domains

There are 37 shared domains, and only 6 misalignment's.

Domain	CoAID Classification	Wikipedia Classification
Mother Jones	FAKE	Generally_reliable
Business Insider	FAKE	No_consensus
The Hill	FAKE	Generally_reliable
The Atlantic	FAKE	Generally_reliable
Newsweek	FAKE	Generally_reliable
Medium	REAL	Generally_unreliable

3.3.5 Results: Information quality

We ran tests to compare the differences in overall classification between our method, and the veracity method. This was run on the classified top and bottom COVID datasets. These each had a size of: Bottom: 6,171, Top: 258,193. The inputting datasets contained (uid user_id_str, cid ’content id’ id_str of the tweet, and url).

Below are the outputs of this computation:

- Bottom debunk had 62: Veracity was (1)
- Bottom disinfo had 78: Veracity was (-1)
- Bottom unreliable had 6,031 (before removing N/a)

Bottom unreliable veracity spread

Veracity score	Number of instances
-1	80
-0.25	836
0	1
0.25	100
0.75	942
1	413

3,655 N/a values removed

59.2286501% of URL's unclassified for Bottom dataset

- Top debunk had 3,654: Veracity was (1)
- Top disinfo had 9110: Veracity was (-1)
- Top unreliable had 245,779 (before removing N/a)

Top unreliable veracity spread:

Veracity score	Number of instances
-1	4571
-0.5	6
-0.333	21
-0.25	15,819
0	740
0.25	7,588
0.75	101,863
1	4,206

110,993 N/a values removed

42.9883847% of URL's unclassified for Top dataset

Some interesting points of that I noticed while investigating the top classifications, was that approximately 60% of the URLs in the veracity of -0.25 were youtube.com. Overall the substantial number of classification made aside from -1 and 1 was promising as it allowed us to gauge a more diverse range of information quality. The rating 1.0, picked up a substantial number of health related links, this was very desirable as particularly in the context of the COVID dataset, many of these links were not included in our initial sourcelist, thus greatly increasing the overall number of classifications.

The rating of 0.75 absolutely dominated the classifications. This was due to us assigning media organisations to this rating. We noted that out of these, there were 22,769 instances of abc.com.au URLs, however there were, zero links to the RMIT fact check subdomain. We also noted that due to the Wikipedia Perrienal classificatins, dailymail.co.uk and thesun.co.uk made up more than 80% of the classified links. Although for Wikipedia's context, they provide a reasonable justification for considering the Daily Mail as disinformation, in our context these links should be reassigned a much lower veracity score.

It was not possible to test linked data together, as the joining system only allowed via the 'uid', which is only unique for each user, and not for each piece of content as 'cid' would.

These are showing the distributions of veracity scores from the COVID TOP DATASET

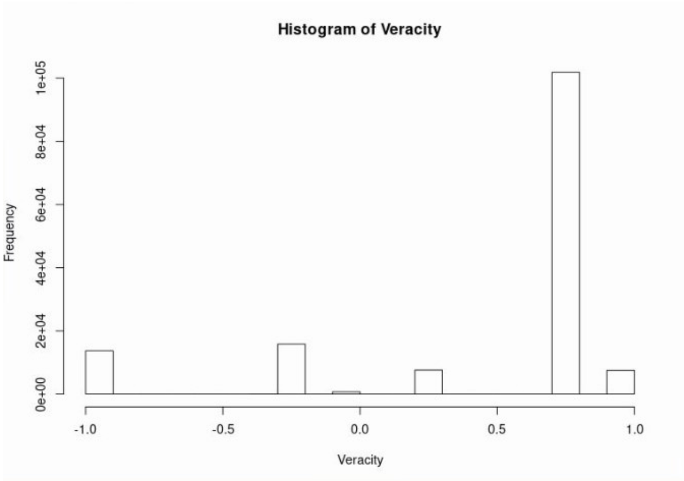


FIGURE 3.5: Histogram of veracity scores for COVID top dataset

Overall, improvements such a utilising the unwound URL element improved our classification rates. However, the approach of using veracity scores to measure the 'reliability' of URL's has been problematic due to the difficulty of classifying the sources of URLs, such as the high proportion of 'thesun' and 'dailymail' links. Furthermore, there is still issue of missing sources, particularly towards the Australian context, (9 News, 7 News, News.com.au, etc) Assessing the reliability of these sources is out of our scope, and their absence from the sourcelists means they are 'useless' to us.

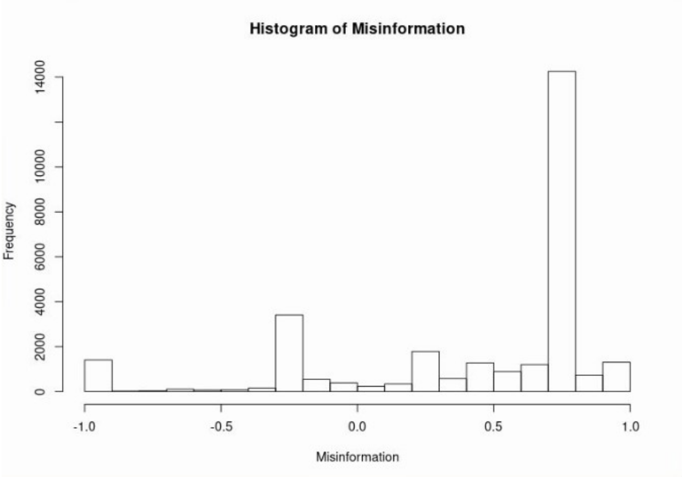


FIGURE 3.6: Histogram of misinformation scores for COVID top dataset

CHAPTER 4

Conclusions

4.1 Implications

The primary contribution of this paper is the new method of classifying URLs. Our method for re-purposing an existing dataset, and applying innovative and creative techniques such as combining veracity measures with existing URL sourcelists allowed us to make classifications out of our limited dataset. Our methods of detecting disinformation tweet spread after URL classification, was supported by findings from other research; and our finding that bottom level debunking has a greater response rate than other classification shows that organic fact checking can be effective.

Due to time constraints, our project had to end before we achieved all of our plans, however we still provide a number of paths to follow up our investigation, and maximise the potential of this methodological approach.

4.2 Limitations

Throughout this project, I greatly struggled with my personal skill level. I had to learn R on the go, and this was my first time dealing with such large datasets. This meant that many of the methods I developed for dealing with the data were inefficient, and I had to constantly re-design and re-develop my processes. A specific example was my trouble with extracting the sub domains (i.e Fox News as opposed to Fox opinion pieces; and RMIT fact check) limiting the overall accuracy of these results.

4.2.0.1 Dataset Limitations:

The dataset not being tailored specifically for this project meant fact checks were few and far between, but overall it meant there was a lack of opportunities for analysis, and we spent much of our time trying to come up with creative solutions to squeeze the most out of the data we had.

4.2.0.2 Sourcelist/URL limitations:

One of the most frustrating limitations, was our ability to classify URL reliability. This is a very complicated issue as there were lots of edge cases, such as links to user driven content hubs such as YouTube/Facebook/Reddit, and there is a massive disparity between reliability of content on these sites. Without investigating each link, and finding a method to classify the reliability of each channel/user on those platforms, which would require a substantially large project, the reliability of those domains had to be classified as 'questionable reliability'.

4.2.0.3 Computational power problems:

The process of filtering through the dataset was very computationally intensive. Although access to the UTS iHPC computing network was provided by my supervisors; the trial-and-error nature of our investigation into the dataset necessitated the re-designing of our filtration processes, and re-running of our datasets. At times this would bottleneck our progress. A specific example would be having to perform separate flirtations for each tweet type, which was an issue that we encountered mid-way through problem. This process took an immense amount of time. Furthermore, in my efforts to try and reduce the size of data the computer had to process, I trimmed the initial tweet data such that only a few necessary JSON fields were included, which we discovered would be required again in the later stages of Problem 2, requiring the data to be re-run and processed again.

4.2.0.4 Limitations of exclusively analysing URLs:

Theoretically instances of ’not good’ fact checks may not include URLs, also there could be cases where the checker makes their claim, and substantiates it later, however these were not included due to the difficulty of doing so.

Also we missed instances of someone making a claim in text, and getting respond to by someone with a URL. Our analysis only picked up instances of URL to URL discussions.

4.3 Future outlook

As this project came to close, there remained a number of unfinished avenues of investigation, that may have enhanced this methodological approach of URL based fact checking analysis.

These plans include:

- Performing an investigation focusing on users -> Investigating official media accounts and seeing if there is a difference in reaction/outreach
- Fully implementing the plans we had for measuring veracity
- Including image elements, and treating them as URLs

4.3.0.1 Measuring pairs of veracity scores

We created a detailed plan to implement a veracity and agreement measure, but were unable to fully implement due to time constraints. Instead we kept with the default veracity classification method outlined by the paper.

Our initial plans for the implementation of a veracity score was based upon this formula:

$$\tilde{\nu} = \frac{(\nu_1 + \nu_2)}{2}$$

This will average the veracity of URl from the top tweet, and that of the bottom tweet.

The results should appear as such: representing (fake,true)

$$\nu\epsilon(-1,1)$$

Formula for measuring Agreement: representing disagreement, agreement

$$a_1 = 1 - \frac{1}{2}|\nu_1 - \nu_2|\epsilon(0,1)$$

This process worked by getting the absolute value of the difference between the veracity of each URL, and using the average to tell the direction (true vs false) To optimise the usage of this method, the measuring the both the veracity and agreement of pairs of URLs would allow identification from similarly aligned, to almost polar opposite instances of fact checking sources, deepening our understandings of what works for different types of sources

4.3.0.2 User properties of fact checking interactions

We initially wanted to investigate the specific differences of organic user created fact checking interactions, and fact checking interactions originating from official accounts. We compiled a list of twitter account names of prominent media organisations, including: “9NewsSyd", "9NewsAUS", "7NewsAustralia", "PerthMediaNews", "BBCNewsAus", "GuardianAus", "cnnbrk", "CNN", "nytimes", "BBCBreaking", "BBCWorld", "TheEconomist", "washington-post", "TIME", "ABC", "ndtv", "AP", "HuffPost", "guardian", "BreakingNews", "SkyNews", "AJEnglish", "FT", "SkyNewsBreak", "politico", "CNBC", "FRANCE24", "guardiannews", "Independent", "BBCAfrica", "Newsweek", "Telegraph”.

We intended to measure the differences between the spread (number of retweet, quote and reply counts) of fact checks originating from these accounts, and those of unclassified (or organic users) user accounts. We did not have the time do perform this analysis, however, we did run some investigations into the proportion of tweets originating from these accounts in our dataset, and the results are as follows:

TABLE 4.1: COVID Dataset

Level	Organic Fact Checks	Official Fact Checks
Top:	255,364	633
Bottom	5809	65

TABLE 4.2: BUSHFIRE DATASET

Level	Organic Fact Checks	Official Fact Checks
Top:	102,559	725
Bottom	2289	12

COVID Dataset

633 instances of top level posting from fact checkers this is 0.2472685227%

255,364 'organic'

65 instance of bottom level posting from fact checkers this is 1.1065713313%

5809 'organic'

BUSHFIRE Dataset

725 instances of top level posting from fact checkers this is 0.7019480268%

102,559 'organic'

12 instance of bottom level posting from fact checkers this is 0.5215123859%

2289 'organic'

4.3.0.3 Other sources of information

The role of additional media objects in Tweets such as images was suggested by my supervisors as a way to utilise more of our dataset, as the spreading of misinformation on Twitter is frequently done through images.

The presence of a media element in the top level tweet was to be treated as a substitute for a URL. Analysis of the images content was not necessary, although could provide greater levels of context.

CHAPTER 5

Figures

Due to the difference in storage for top-level and bottom-level tweets, separate graphs were produced for each tweet type, and count type.

Note: These represent Top level tweets (Original tweets) and Bottom level tweets (responding tweets) that only contain URLs

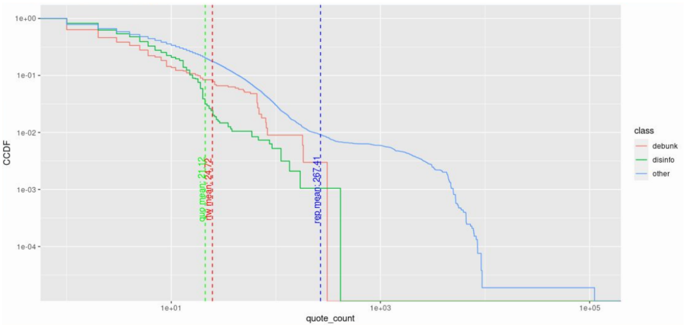


FIGURE 5.1: ECDF Plot of Bushfire Bottom-Level Quote Counts

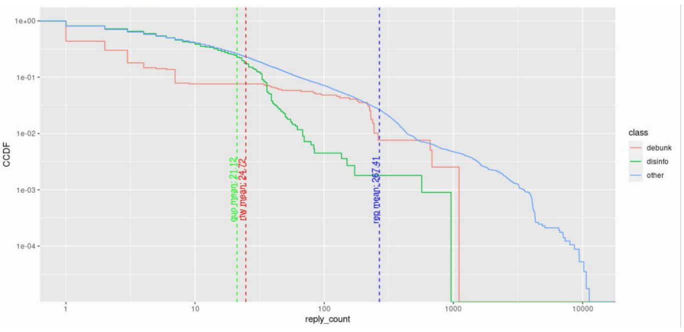


FIGURE 5.2: ECDF Plot of Bushfire Bottom-Level Reply Counts

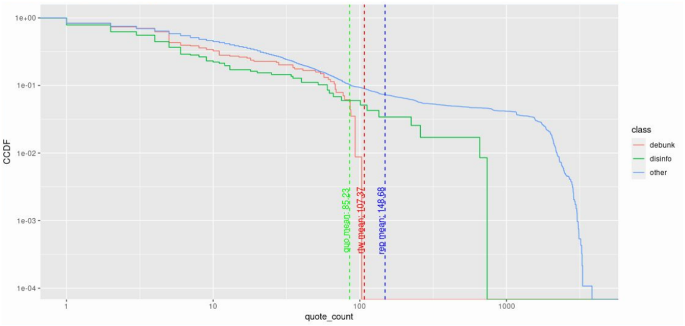


FIGURE 5.3: ECDF Plot of Bushfire Top-Level Quote Counts

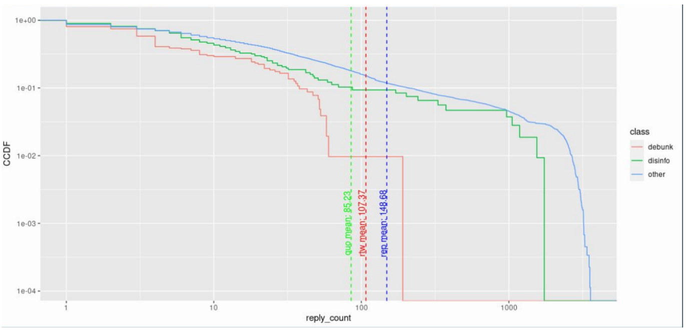


FIGURE 5.4: ECDF Plot of Bushfire Top-Level Reply Counts

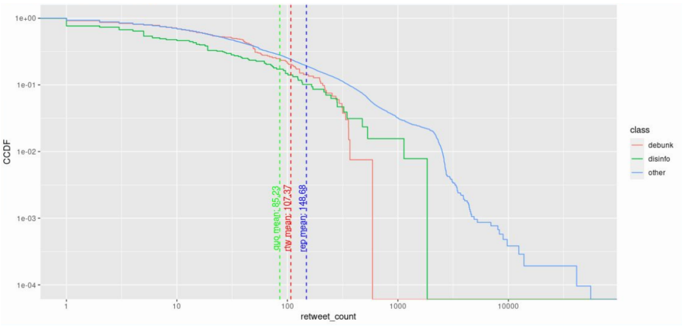


FIGURE 5.5: ECDF Plot of Bushfire Top-Level Retweet Counts

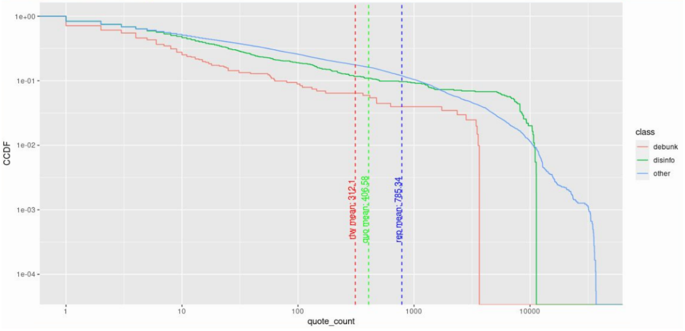


FIGURE 5.6: ECDF Plot of COVID Top-Level Quote Counts

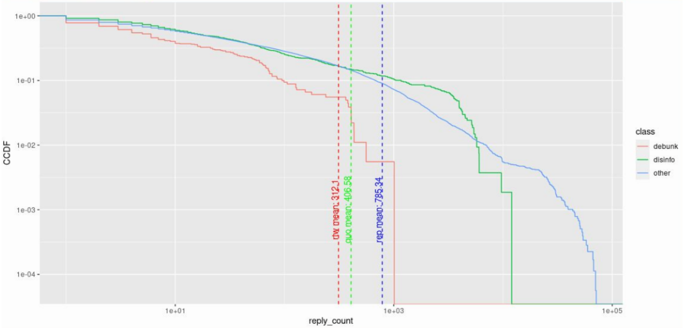


FIGURE 5.7: ECDF Plot of COVID Top-Level Reply Counts

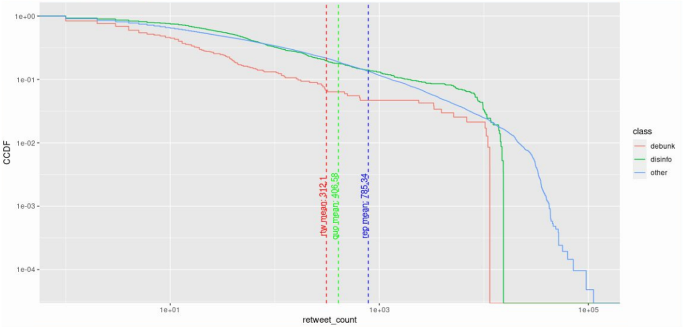


FIGURE 5.8: ECDF Plot of COVID Top-Level Retweet Counts

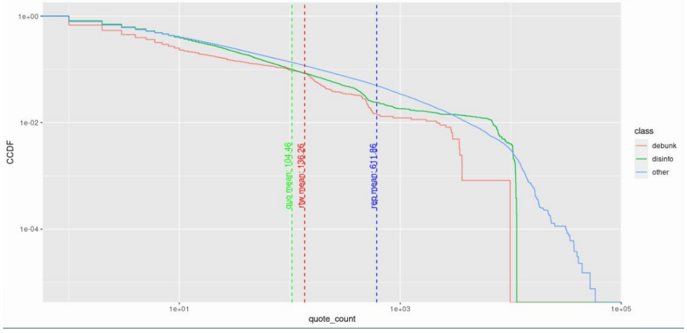


FIGURE 5.9: ECDF Plot of COVID Bottom-Level Quote Counts

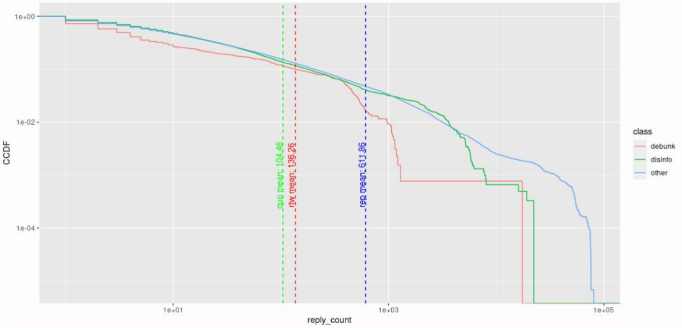


FIGURE 5.10: ECDF Plot of COVID Bottom-Level Reply Counts

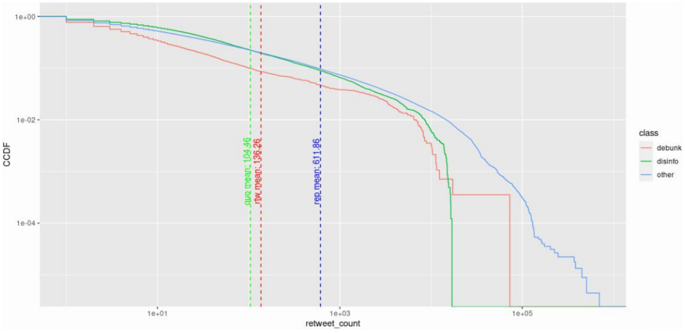


FIGURE 5.11: ECDF Plot of COVID Bottom-Level Retweet Counts

Bibliography

[1] K.S. Aboufarw, A. Grigorev and A.S. Mihaita. 'Traffic accident risk forecasting using Vision Transformers,' in: *Proc. of the IEEE Intelligent Transport Systems Conference 2022, Macao, China*. 2022.

[2] Anonymous. *Facebook whistleblower: Internal documents detail how misinformation spreads to users - CBS News*. Oct. 2021. URL: https://www.cbsnews.com/news/facebook-whistleblower-frances-haugen-documents-misinformation-spread/.

[3] Nik Dawson et al. 'Adaptively selecting occupations to detect skill shortages from online job ads'. In: *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*. Los Angeles, CA, USA: IEEE, Dec. 2019, pp. 1637–1643. ISBN: 9781728108582. DOI: 10.1109/BigData47090.2019.9005967. arXiv: 1911.02302. URL: http://arxiv.org/abs/1911.02302%20https://ieeexplore.ieee.org/document/9005967/.

[4] Nikolas Dawson et al. 'Layoffs, inequity and COVID-19: A longitudinal study of the journalism jobs crisis in Australia from 2012 to 2020'. In: *Journalism* (Feb. 2021), p. 146488492199628. ISSN: 17413001. DOI: 10.1177/1464884921996286. arXiv: 2008.12459. URL: https://arxiv.org/abs/2008.12459%20http://journals.sagepub.com/doi/10.1177/1464884921996286.

[5] Adrien Friggeri et al. 'Rumor Cascades'. In: *Proceedings of the International AAAI Conference on Web and Social Media* 8 (1 May 2014), pp. 101–110. ISSN: 2334-0770. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14559.

[6] Keren Goldshlager and Aaron Berman. *Facebook Announces New Ratings for Fact-Checking Partners*. Aug. 2020. URL: https://www.facebook.com/journalismproject/programs/third-party-fact-checking/new-ratings.

[7] A. Grigorev et al. 'Traffic incident duration prediction via a deep learning framework for text description encoding'. In: *Proc. of the IEEE Intelligent Transport Systems Conference 2022, Macao, China*. 2022.

[8] Artur Grigorev et al. 'Incident duration prediction using a bi-level machine learning framework with outlier removal and intra–extra joint optimisation'. In: *Transportation Research Part C: Emerging Technologies* 141 (2022), p. 103721. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2022.103721. URL: https://www.sciencedirect.com/science/article/pii/S0968090X22001589.

[9] Mirya R. Holman and J. Celeste Lay. 'They See Dead People (Voting): Correcting Misperceptions about Voter Fraud in the 2016 U.S. Presidential Election'. In: *https://doi.org/10.1080/15377857.2018.1478656* 18 (1-2 Apr. 2018), pp. 31–68. ISSN: 15377865. DOI: 10.1080/15377857.2018.1478656. URL: https://www.tandfonline.com/doi/abs/10.1080/15377857.2018.1478656.

[10] F. Issa et al. 'An Intelligent System based on Natural Language Processing to support the brain purge in the creativity process'. In: *IAENG International Conference on Artificial Intelligence and Applications (ICAIA'14) Hong Kong* (Mar. 2014).

[11] Quyu Kong, Marian Andrei Rizoiu and Lexing Xie. 'Describing and Predicting Online Items with Reshare Cascades via Dual Mixture Self-exciting Processes'. In: *International Conference on Information and Knowledge Management, Proceedings*. New York, NY, USA: ACM, Oct. 2020, pp. 645–654. ISBN: 9781450368599. DOI: 10.1145/3340531.3411861. arXiv: 2001.11132. URL: https://arxiv.org/pdf/2001.11132.pdf%20https://dl.acm.org/doi/10.1145/3340531.3411861.

[12] Quyu Kong et al. 'Mapping Online Problematic Content using Ethnographic and Qualitative Analysis augmented with Human-in-the-loop Machine Learning'. In: (2021). URL: www.aaai.org.

[13] Quyu Kong et al. ‘Will This Video Go Viral: Explaining and Predicting the Popularity of Youtube Videos’. In: *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018*. Lyon, France: ACM Press, 2018, pp. 175–178. ISBN: 9781450356404. DOI: 10.1145/3184558.3186972. arXiv: 1801.04117. URL: https://arxiv.org/abs/1801.04117%20http://dl.acm.org/citation.cfm?doid=3184558.3186972.

[14] T. Mao, A.S. Mihaita and C. Cai. ‘Traffic Signal Control Optimisation under Severe Incident Conditions using Genetic Algorithm’. In: *Proc. of ITS World Congress (ITSWC 2019), Singapore* (Oct. 2019).

[15] Tuo Mao et al. ‘Boosted Genetic Algorithm Using Machine Learning for Traffic Control Optimization’. In: *Trans. Intell. Transport. Sys.* 23.7 (July 2022), pp. 7112–7141. ISSN: 1524-9050. DOI: 10.1109/TITS.2021.3066958. URL: https://doi.org/10.1109/TITS.2021.3066958.

[16] Drew B. Margolin, Aniko Hannak and Ingmar Weber. ‘Political Fact-Checking on Twitter: When Do Corrections Have an Effect?’ In: *https://doi.org/10.1080/10584609.2017.1334018* 35 (2 Apr. 2017), pp. 196–219. ISSN: 10917675. DOI: 10.1080/10584609.2017.1334018. URL: https://www.tandfonline.com/doi/abs/10.1080/10584609.2017.1334018.

[17] Paul Mena. ‘Cleaning Up Social Media: The Effect of Warning Labels on Likelihood of Sharing False News on Facebook’. In: *Policy & Internet* 12 (2 June 2020), pp. 165–183. ISSN: 1944-2866. DOI: 10.1002/POI3.214. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/poi3.214%20https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.214%20https://onlinelibrary.wiley.com/doi/10.1002/poi3.214.

[18] A. S. Mihaita et al. ‘Air quality monitoring using stationary versus mobile sensing units: a case study from Lorraine, France’. In: *Proc. of ITS World Congress (ITSWC 2018), Copenhagen, Denmark* (Sept. 2018).

[19] A. S. Mihaita et al. ‘Predicting air quality by integrating a mesoscopic traffic simulation model and air pollutant estimation models’. In: *International Journal of Intelligent Transportation System Research (IJITSR)* 17.2 (2019), pp. 125–141. ISSN:

1868-8659. DOI: DOI:10.1007/s13177-018-0160-z. URL: https://link.springer.com/article/10.1007/s13177-018-0160-z.

[20] A.S. Mihaita, H. LI and M.A. Rizoiu. *Traffic congestion anomaly detection and prediction using deep learning*. 2020. DOI: arXiv:2006.13215.

[21] Adriana-Simona Mihaita, Zac Papachatgis and Marian-Andrei Rizoiu. ‘Graph Modelling Approaches for Motorway Traffic Flow Prediction’. In: *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. Rhodes: IEEE Press, 2020, pp. 1–8. DOI: 10.1109/ITSC45102.2020.9294744. URL: https://doi.org/10.1109/ITSC45102.2020.9294744.

[22] Adriana-Simona Mihaita et al. ‘Arterial incident duration prediction using a bi-level framework of extreme gradient-tree boosting’. In: *Proceedings of the 26th ITS World Congress*. Singapore, May 2019, pp. 1–12. arXiv: 1905.12254. URL: http://arxiv.org/abs/1905.12254.

[23] Adriana-Simona Mihaita et al. ‘Motorway Traffic Flow Prediction using Advanced Deep Learning’. In: *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. Auckland, New Zealand: IEEE, Oct. 2019, pp. 1683–1690. ISBN: 978-1-5386-7024-8. DOI: 10.1109/ITSC.2019.8916852. arXiv: 1907.06356. URL: https://ieeexplore.ieee.org/document/8916852/.

[24] A.S. Mihăiţă, M. Camargo and P. Lhoste. ‘Evaluating the impact of the traffic reconfiguration of a complex urban intersection’. In: *10th International Conference on Modelling, Optimization and Simulation (MOSIM 2014), Nancy, France, 5-7 November 2014* (Nov. 2014).

[25] Adriana Simona Mihăiţă et al. ‘An investigation of positioning accuracy transmitted by connected heavy vehicles using DSRC’. In: *Transportation Research Board - 96th Annual Meeting, Washington, D.C.* (Jan. 2017).

[26] Simona Mihăiţă and Stéphane Mocanu. ‘An energy model for event-based control of a switched integrator’. In: *IFAC Proceedings Volumes* 44.1 (2011). 18th IFAC World Congress, pp. 2413–2418. ISSN: 1474-6670. DOI: https://doi.org/10.3182/20110828-6-IT-1002.02082. URL: https://www.sciencedirect.com/science/article/pii/S1474667016439741.

[27] Swapnil Mishra, Marian-Andrei Rizoiu and Lexing Xie. 'Modeling Popularity in Asynchronous Social Media Streams with Recurrent Neural Networks'. In: *International AAAI Conference on Web and Social Media (ICWSM '18)*. Stanford, CA, USA, 2018, pp. 1–10. URL: https://arxiv.org/pdf/1804.02101.pdf.

[28] D. Monticolo and A.S. Mihăiţă. 'A multi Agent System to Manage Ideas during Collaborative Creativity Workshops'. In: *International Journal of Future Computer and Communication (IJFCC)* 3.1 (Feb. 2014), pp. 66–70. ISSN: 2010-3751. DOI: 10.7763/IJFCC.2014.V3.269.

[29] Nick Pickles and Yoel Roth. *Updating our approach to misleading information*. May 2020. URL: https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.

[30] Marian Andrei Rizoiu and Julien Velcin. 'Topic extraction for ontology learning'. In: *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances*. Ed. by Wilson Wong, Wei Liu and Mohammed Bennamoun. IGI Global, 2011, pp. 38–60. ISBN: 9781609606251. DOI: 10.4018/978-1-60960-625-1.ch003. URL: http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-60960-625-1.ch003.

[31] Marian Andrei Rizoiu et al. 'Evolution of privacy loss in Wikipedia'. In: *WSDM 2016 - Proceedings of the 9th ACM International Conference on Web Search and Data Mining*. ACM. New York, New York, USA: ACM Press, Dec. 2016, pp. 215–224. ISBN: 9781450337168. DOI: 10.1145/2835776.2835798. arXiv: 1512.03523. URL: http://dl.acm.org/citation.cfm?doid=2835776.2835798%20http://arxiv.org/abs/1512.03523%20http://dx.doi.org/10.1145/2835776.2835798.

[32] Marian-Andrei Rizoiu and Lexing Xie. 'Online Popularity under Promotion: Viral Potential, Forecasting, and the Economics of Time'. In: *International AAAI Conference on Web and Social Media (ICWSM '17)*. Montréal, Québec, Canada, 2017, pp. 182–191. URL: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15553%20https://arxiv.org/pdf/1703.01012.pdf.

[33] S. Shaffiei, A.S. Mihaita and C. Cai. ‘Demand Estimation and Prediction for Short-term Traffic Forecasting in Existence of Non-recurrent Incidents’. In: *Proc. of ITS World Congress (ITSWC 2019), Singapore* (Oct. 2019).

[34] Sajjad Shafiei et al. ‘Integrating data-driven and simulation models to predict traffic state affected by road incidents’. In: *Transportation Letters* 14.6 (2022), pp. 629–639. DOI: 10.1080/19427867.2021.1916284. eprint: https://doi.org/10.1080/19427867.2021.1916284. URL: https://doi.org/10.1080/19427867.2021.1916284.

[35] Sajjad Shafiei et al. ‘Short-Term Traffic Prediction under Non-Recurrent Incident Conditions Integrating Data-Driven Models and Traffic Simulation’. In: *Transportation Research Board (TRB) 99th Annual Meeting, Washington D.C.* 2020. DOI: http://hdl.handle.net/10453/138721.

[36] Juliette T.H. Unwin et al. ‘Using hawkes processes to model imported and local malaria cases in near-elimination settings’. In: *PLoS Computational Biology* 17.4 (Apr. 2021). Ed. by Alex Perkins, e1008830. ISSN: 15537358. DOI: 10.1371/JOURNAL.PCBI.1008830. URL: http://medrxiv.org/content/early/2020/07/17/2020.07.17.20156174.abstract%20https://dx.plos.org/10.1371/journal.pcbi.1008830.

[37] Soroush Vosoughi, Deb Roy and Sinan Aral. ‘The spread of true and false news online’. In: *Science* 359 (6380 Mar. 2018), pp. 1146–1151. DOI: 10.1126/SCIENCE.AAP9559.

[38] Nathan Walter et al. ‘Political Communication Fact-Checking: A Meta-Analysis of What Works and for Whom’. In: (2019). ISSN: 1091-7675. DOI: 10.1080/10584609.2019.1668894. URL: https://www.tandfonline.com/action/journalInformatio journalCode=upcp20.

[39] Tao Wen et al. ‘Integrated Incident Decision-Support using Traffic Simulation and Data-Driven Models’. In: *Transportation Research Record* 2672.42 (2018), pp. 247–256. DOI: 10.1177/0361198118782270. eprint: https://doi.org/10.1177/0361198118782270. URL: https://doi.org/10.1177/0361198118782270.

[40] Siqi Wu, Marian Andrei Rizoiu and Lexing Xie. ‘Variation across scales: Measurement fidelity under Twitter data sampling’. In: *Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020*. Mar. 2020, pp. 715–725. ISBN: 9781577357889. arXiv: 2003.09557. URL: https://arxiv.org/abs/2003.09557.

[41] Siqi Wu, Marian-Andrei Rizoiu and Lexing Xie. ‘Estimating Attention Flow in Online Video Networks’. In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (Nov. 2019), pp. 1–25. ISSN: 25730142. DOI: 10.1145/3359285. URL: http://dl.acm.org/citation.cfm?doid=3371885.3359285.

[42] Rui Zhang, Christian Walder and Marian-Andrei Rizoiu. ‘Variational Inference for Sparse Gaussian Process Modulated Hawkes Process’. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.04 (Apr. 2020), pp. 6803–6810. ISSN: 2374-3468. DOI: 10.1609/aaai.v34i04.6160. arXiv: 1905.10496. URL: http://arxiv.org/abs/1905.10496%20https://aaai.org/ojs/index.php/AAAI/article/view/6160.

[43] D. Zhao et al. ‘Real-time attention-augumented spatio-temporal networks for video-based driver activity recognition’. In: *Proc. of the IEEE Intelligent Transport Systems Conference 2022, Macao, China*. 2022.

1 Appendix

Wikipedia Perennial Sourcelist: https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources/Perennial_sources

External Fact Checking Sourcelist: https://ifcncodeofprinciples.poynter.org/signatories

urltools Package: https://cran.r-project.org/package=urltools

CoAID Dataset: https://github.com/cuilimeng/CoAID

Data Dictionary for Tweets in the v1 API: https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/entities