D4.2 – Crowdsourced information clustering

Grant agreement number:	101003606	Due date of Deliverable:	[30 Sept 2021]
Start date of the project:	1 April 2020	Actual submission date:	[30 Sept 2021]
Duration:	36 months	Deliverable approved by the WPL/CO : ☐	

Lead Beneficiary: The Open University (OU)

Contributing beneficiaries: SVENSKA HANDELSHOGSKOLAN (HAN), TECHNISCHE UNIVERSITEIT DELFT (TUD), CENTRUM BADAN KOSMICZNYCH POLSKIEJ AKADEMII NAUK (CBK)

Keywords
COVID19, misinformation, fact checking, classification, social media

Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

History			
Author	**Date**	**Reason for change**	**Release**
Harith Alani, Grégoire Burel and Tracie Farrell	20/08/2021	Structure and outline	V1
Harith Alani, Grégoire Burel and Tracie Farrell	16/09/2021	Draft 1 all chapters	V2
Anna Foks-Ryznar	24/09/2021	Contribution to chapter 2	V2.4
Kees Boersma	24/09/2021	Review	V2.5
Harith Alani, Grégoire Burel, Tracie Farrell and Henry Lachman and Anna Foks-Ryznar	28/09/2021	Post-Review by Kees Boersma	V3


HEROS

Executive Summary

This deliverable deals with the task of automatically processing crowd information to enhance the situational awareness of citizens during a crisis: how the crisis is unfolding on the ground, what needs are requested by the public, and what genre or (mis)information is shared by the public on online social media. We use the model of mutual aid to understand the motivations that citizens may have in sharing material goods, their time, or even information. We show how our Fact-Checking observatory allows us to track misinformation about COVID-19 on Twitter, using annotations on the credibility of information (provided by the fact-checking community). We demonstrate how the data used in the observatory can be further repurposed to aid in the classification of new information, based on its topic. We also show how Twitter users' biolines can aid in understanding more about which (mis)information is interesting to which social or cultural groups, through following hashtag occurrences in misinforming and fact-checking posts. Finally, we introduce our future plans to use the tools we have developed to explore the information sharing behaviours of Twitter users interested in providing mutual aid during COVID-19.

Table of content

Introduction **7**
Objectives 8
Relationships to other work packages 8
Contribution 9
Structure of Document 9

Temporal Characteristics of Misinformation **9**
Fact-Checking Observatory 10
Data Collection 10
Weekly Reports 11
Design Update 14
Web mapping application 16

Classifying COVID-19 posts on social media **17**
Topic Extraction and Classification 17
User characteristics and COVID-19 18

Covid-19 (Mis)information Topic Classification **19**
Proxied Twitter Covid-19 Topic Classification 19
Dataset and Data Collection 20
Fact-check URLs and topic dataset 21
Twitter dataset 22
Topic Classification Models 22
Database Post-processing and Dataset generation 22
Model Training and Evaluation 23
Classification Results 24
Qualitative Error Analysis 26
General explanations 26
Topic Specific Explanations 28
Authorities 28
Causes 28
Conspiracy theory 29
Spread 29
Cures 30
Vaccine 30
Discussion and Future Work 31

User Descriptions Co-Occurrence Hashtags Analysis **32**
Data Collection and Co-Occurrences Generation 32
Co-Occurrences: Further Qualitative Analysis and Future Work 37

Crowd-sourcing Mutual Aid **39**

© HERoS Consortium

Planned Research 41

Conclusion **43**

References **43**

Table of tables

Table 1: Classification evaluation results for different models. 25
Table 2: Per class classification results for the Distil-BERT model. 26

Table of figures

Figure 1: Automated Tweets of new reports by the Fact-Checking Observatory 14
Figure 2: Home page of FCO, showing available recent reports. 15
Figure 3: Amount of misinformation topics shared over time on Twitter. 17
Figure 4: Relative quantity of misinforming claims written in different languages. 18
Figure 5: Misinformation and Fact-checking spread across different demographics.
Top: Gender, Center: Age group, Bottom: Account type. 18
Figure 6: Fact-Checking Observatory new design. 19
Figure 7: Interactive maps of misinformation (left) and fact-checks occurrences (right).
Top: number of occurrences, Bottom: the most recent top topic. 20
Figure 8: Proportion of posts with a given topic for the final train, development and test datasets. 27
Figure 9: Confusion matrix for misclassified posts 29
Figure 10: Hashtag Co-occurrences of users sharing misinformation 39
Figure 11: Hashtag Co-occurrences of users sharing fact-checks 40
Figure 12: Hashtag Co-occurrences of users sharing misinformation 41
Figure 13: Hashtag Co-occurrences of users sharing misinformation 42
Figure 14: Top hashtag categories of users sharing misinformation 43

© HERoS Consortium

1 Introduction

Being able to automatically and accurately process crowd information is important for enhancing the situational awareness of citizens (regardless of public role) on how a crisis is unfolding on the ground, what needs are requested by the public, and what genre or (mis)information is shared by the public on online social media. In this deliverable we describe our efforts to identify sources of data from which we can gather insights about what kinds of needs (informational and tangible) citizens have shared during the COVID-19 pandemic. We introduce our methodological approaches for processing this data and making it a useful resource for the research community, as well as government officials and authorities who are charged with handling the crisis and understanding the nature of citizen involvement or partnership. In particular, we are interested in tracking and amplifying the efforts of citizens to meet each other's needs during the crisis. We focus on tangible needs, tackled by local mutual aid groups and volunteers responding to government guidance, and the informational needs, tackled by fact-checkers and citizens who share fact-checks and corrective information on social media. Our research aims to explore why individuals might provide assistance during a crisis, whether online or offline, and how we can understand and support their actions at scale.

The term mutual aid comes from anarchist philosopher, Peter Kropotkin[1], with the focus on cooperation - how we work together to solve societal problems. Kropotkin argued that human society does not have individualistic aims that drive evolution, as Darwin might have proposed. Rather, he believed that cooperation benefits us as a whole. Moreover, we can see this in how we choose to come together in a crisis. As we will describe below, COVID-19 has been an excellent example of this type of cooperation, with enormous efforts happening both online and offline to provide verified information and service to one another, so that we can navigate this challenge as a society.

From our previous research on “misinformation resilience”[2], we became aware of ways in which citizens access information, how they understand credibility and make decisions about what to share on social media. We are encouraged that social media users do share information with one another (for a variety of reasons) (Amazeen et al., 2019), but we still need to understand a lot more about how to present factual information and how to encourage sharing verified information whenever possible. This deliverable describes our current research on understanding user characteristics that may assist with that goal.

In addition, we learned that fact-checking has become standardised, with the creation of the International Fact Checking Network[3] (IFCN) by Poynter and standardisation of practices around the ClaimReview schema[4]. Fact-checkers are now accredited by the IFCN and their work is aggregated on the network's website, structured through the ClaimReview schema such that it is possible to gather data about who checked which information, at what time, from which location and what the fact-checker determined about the truthfulness of the information. We developed our Fact-checking Observatory, which we discuss in this deliverable, using this data. The observatory allows us to track the spread of misinformation and

[1]Kropotkin’s concept of mutual aid, http://www.bzby.dk/tankekriminalitet/bib/Mutual.Aid-Kropotkin.pdf
[2] European Horizon 2020 project Co-Inform, https://coinform.eu/
[3] International Fact-checking Network from Poynter, https://www.poynter.org/ifcn/
[4] Claim Review Schema, https://www.claimreviewproject.com/

fact-checks on Twitter, by topic over time. We also include some geographic and demographic information about users on Twitter who point to misinformation and fact-checks through URLs. In this deliverable, we describe how we plan to enhance this work by automatically classifying new information and connecting more detailed user characteristics to accompany our analysis of topics and their spread over time. We report on some of our early experiments on training a classifier using the IFCN data from Poynter, and a qualitative analysis of errors we identified during our evaluation.

Finally, we also examine the work of mutual aid groups, and incorporate evidence from other research teams on the provision of items and services by volunteers across the COVID-19 pandemic. We outline our approach to studying the work of these groups at scale, using some of the same techniques we employ above for tracking the activities of mutual aid groups during COVID-19.

Objectives

The objectives of this deliverable are as follows:

- to collate the relevant literature on topic classification of, particularly, COVID-19 data on social media
- to demonstrate how automatic classification might be achieved using the crowd-sourced data from fact-checkers in the IFCN
- to explore ways of extending this work to understand more about social media users and their contributions to amplifying this work through their own social media channels and in their own networks
- to describe the activities of mutual aid groups during COVID-19, which may be important to understand at scale
- to outline our approach for studying the behaviour of such groups online and at scale
- to propose some immediate next steps for future research in this area

Relationships to other work packages

Work Package 4 on social media analytics is intended to absorb relevant knowledge from governance in WP1, epidemiological modelling from WP2 and supply chain management from WP3 to model the data we are seeing online. The concepts of mutual aid and social capital come to us from our colleagues at the Vrije Universiteit Amsterdam (VU) who have already been examining spontaneous volunteering and grassroots movements in Amsterdam during the refugee crisis. For our analysis of online misinformation and fact-checking spread, we have been working with our colleagues at CBK, the Space Research Centre of Polish Academy of Sciences, who are transforming the data from our Fact-checking Observatory into interactive geoinformatic maps. Of course, our colleagues in WP3, Technical University Delft, have shared with us various ways to model crisis management responses so that we can make sense of the overall impact of the activities described above on the pandemic. Positions between these work packages that involve tracking and making sense of the pandemic, social media data is able to use and support those technologies.

Contribution

This deliverable is intended to demonstrate new ways of thinking about crowd-sourced social media data during COVID-19, and how this data can support the sense-making activities of citizens during a crisis, as well as some of their behaviours in response. This deliverable makes the following contributions:

- A model for classifying misinforming and fact-checking posts related to COVID-19 by topic over time, using data crowd-sourced from fact-checkers about misinformation circulating on Twitter.
- A methodology for analysing hashtags in user biolines on Twitter to connect user characteristics with the topics of misinformation or fact-checks they share over time. This will allow us to "crowd-source" the annotation of misinformation and fact-checks by user interests and ideologies, making it easier to identify potential topics where perceptual bias may be triggered.
- Application of the theoretical model of mutual aid as a potential way to understand the motivations of citizens to exchange information, time and material goods during a crisis.

Structure of Document

This document is structured as follows:

In section 2, following this introduction, we showcase our Fact-checking Observatory (FCO) and what we have managed to achieve in terms of understanding the temporal flow of misinformation by topic, using the data from Poynter. This includes the development of a web mapping application to provide geolocated data. In section 3, we review some of the relevant literature on classification tasks on COVID-19 data, with a focus on social media data. We then describe our experiments with using the Poynter data to train a new classifier, capable of classifying new social media posts by COVID19-misinformation topics (e.g., vaccine, cure, origin of virus). We provide a qualitative analysis of the algorithm's errors, with some suggestions for future improvements. Finally, we propose using the hashtags from Twitter users' biolines to provide more information on what kinds of users spread information about which topics. We demonstrate the approach with a small experiment on the co-occurence of hashtags to identify user clusters, and provide some direction for future research. In section 4, we review relevant literature on mutual aid groups during COVID-19 and share details from a study we conducted on one local mutual aid group in the UK. In addition, we introduce our approach for establishing and potentially tracking citizen needs through posts on social media. The document concludes with a summary of future research and conclusions about the relevance of tracking social media data to improve our understanding of the COVID-19 pandemic.

2 Temporal Characteristics of Misinformation

Much research has been done on the analysis of the dynamics of misinformation on social media platforms. However, there is still a lack of understanding of how these dynamics change over time, and how different types of misinformation spread over different periods of time. In HERoS, we constructed a collector of COVID19 related misinformation that captures the latest information and updates our database automatically and regularly. In this section, we describe (a) how we can track the spread of COVID19 misinformation and corresponding fact-checks over time, (b) the Fact-Checking Observatory that we developed specifically for tracking and reporting on how misinformation and fact-checks are spreading on Twitter on weekly basis, and (c) the visualisation of spatial and temporal variability of misinformation and fact-checks occurrences.

Fact-Checking Observatory

In D4.1, we gave the first report of the Fact-Checking Observatory (FCO)[5] that was developed in this project with additional seeding support from the UK Higher Education Innovation Fund (HEIF)[6]. FCO is a website for automatically generating reports about the co-spread of COVID-19 misinformation and fact-checks on Twitter.

The FCO tracks the appearance of COVID19 misinformation and corresponding fact-checks on Twitter, and automatically produces reports every week with an up-to-date description and graphs on the topical and demographic spread of misinformation and fact-checks on Twitter. In addition, a dedicated Twitter account for FCO posts about new reports on a weekly basis (https://twitter.com/fc_observatory - Figure 1).

Figure 1: Automated Tweets of new reports by the Fact-Checking Observatory

Data Collection

The data used by the observatory consists of COVID-19 misinformation URLs and their corresponding fact-checks and types as provided by the Poynter CoronaVirusFacts Alliance.[7] They produce a database of such information, collected from more than 70 countries and in over 40 languages.[8] The FCO regularly searches for these URLs on Twitter, and tracks their appearance and disappearance, which is then reported in a suite of tables and graphs to help monitor their change and progress. By 13th September

[5] The Fact-checking Observatory, https://fcobservatory.org.
[6] UK HEIF, https://re.ukri.org/knowledge-exchange/the-higher-education-innovation-fund-heif.
[7] https://www.poynter.org/coronavirusfactsalliance/
[8] https://www.poynter.org/ifcn-covid-19-misinformation/

2021, the FCO database contained 14,420 fact-check URLs and over 487,000 tweets with one of their URLs.

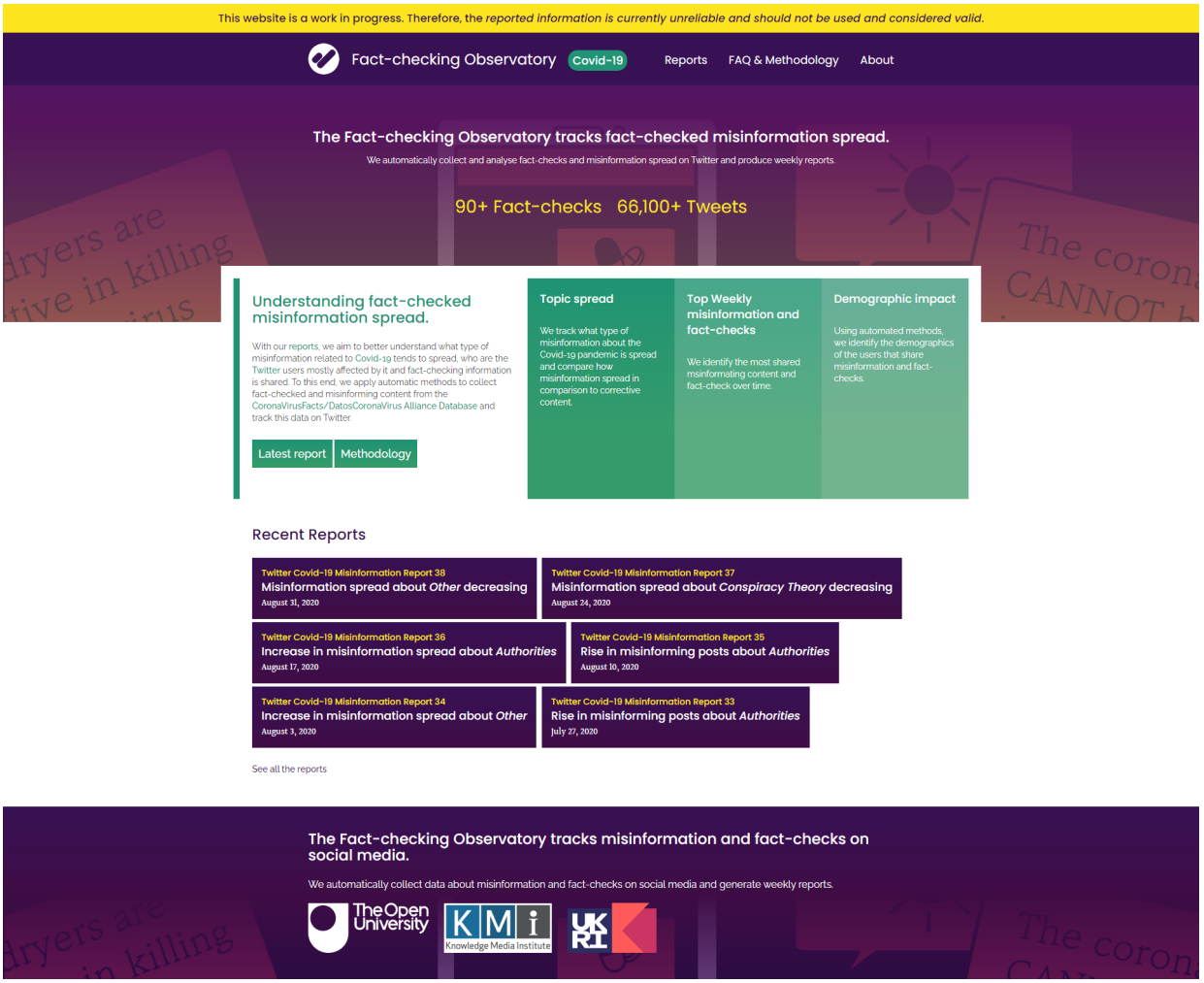

Figure 2: Home page of FCO, showing available recent reports.

Weekly Reports

A key goal of the FCO is to create weekly human readable reports that are not created using any manual input. In order to generate the reports, we use templates that are filled every week as new data is collected. For transparency purposes, templates and data collection are versioned so readers are aware when new data and new templates are used for a given report. In order to not have reports always looking exactly the same, some parts of the reports are generated using different variations of the same sentence at random. For example, the title of a given report may be “Misinformation spread about Authorities decreasing” or “Misinforming posts about Authorities reducing”.

Each report shows the amount of misinformation and fact-check spreading in a given week compared to previous periods with insights about fact-checking organisations, topical spread and demographics. Reports are divided in 5 different sections summarising the evolution of misinformation compared to the previous report and state of the Covid-19 misinformation and fact-checking spread:

1. *Title and subtitle*: The first part of the report contains the title and subtitle that is automatically generated at random based on what topic spread increased or decreased the most during the report period compared to the previous report. This section also includes the version of the report and data used for generating the web page so that readers are aware if a report or underlying data has changed.
2. *Summary statistics and disclaimer*: The second section contains general statistics concerning the data used for the report as well as information about how the report is generated. The summary section gives details about the amount of misinforming tweets collected so far, the number of additional misinformation collected since the previous report and the difference between the current report spread increase and the previous report spread. Similar statistics are also reported for the amount of shared fact-checking URLs as well as the amount of fact-checking URLs used for generating the data so far and the number of fact-checking organisations involved. This section gives a glance about how misinformation or fact-checks spread changes compared to the previous reporting period.
3. *Key content and topics* (Figure Y): In this section, the amount of posts shared over time about covid-related topics and statistics about the topics that were the most and least shared during the reporting period are given for quickly identifying key topics and posts during the reporting period. For instance, this information can be used by fact-checkers to identify what misinformation is spreading the most and plans their response accordingly.
4. *Fact-checking* (Figure Z): This section reports about who is providing the fact-checks such as the language of the fact-checking organisation and their location. Statistics about what topic spreads the most for both misinformation and fact-checks are also reported with details about the current and last report as well as the period before the last two reports.
5. *Demographic impact* (Figure XZ): The last section is based on the automatic extraction of demographic information (Wang et al., 2019) about who shares misinformation and fact-checks with details about which gender, age group and account type spreads misinformation and fact-checks the most. This can be used for better directing fact-checking efforts towards particular demographics or identifying what demographic is the most sensitive to misinformation.

These reports are aimed to help (a) monitor rise/fall in sharing false claims and their corrections, (b) identify the most popular and persistent claims, and (c) assess impact of fact-checks in halting the spread of specific claims.

Reports are generated using the blogdown R package[9] so that templates can incorporate both the fixed textual content necessary for generating the reports and the code necessary to compute the different statistics and generate the different graphs used in the reports.

[9] Blogdown, https://github.com/rstudio/blogdown.

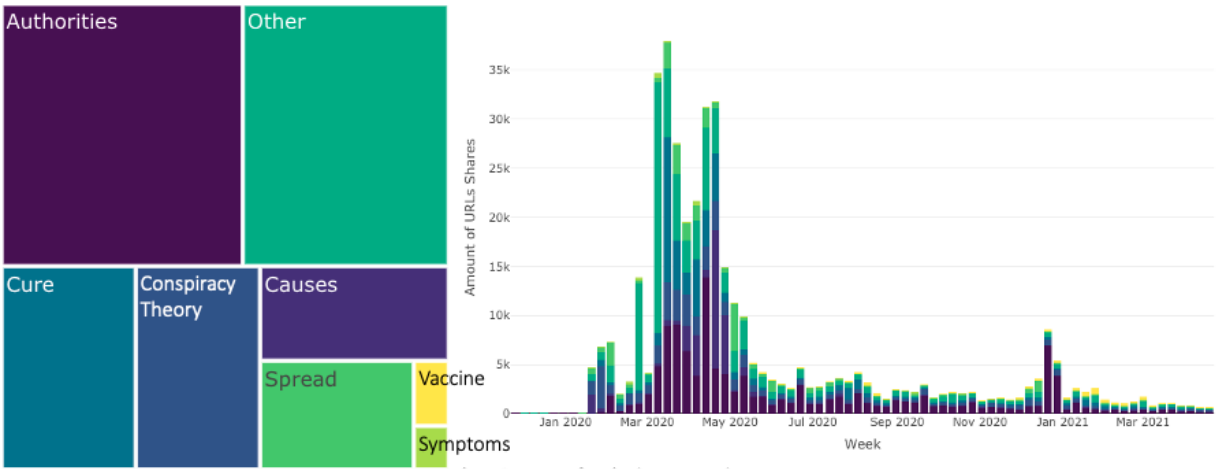

Figure 3: Amount of misinformation topics shared over time on Twitter.

Misinforming URL	Fact-check URL	Topic	Current Week	Previous Week	Total
https://twitter.com/EmeraldRobinson/status/1423274794343190531	FactCheck.org	Vaccine	814	0	814
https://www.youtube.com/watch?v=Du2wm5nhTXY	PolitiFact	Vaccine	229	226	5060
https://www.worldometers.info/	Agencia Ocote	Authorities	171	185	34283
https://madisonarealymesupportgroup.com/2020/09/30/proof-that-the-pandemic-was-planned-with-purpose/	Newschecker	Conspiracy Theory	38	22	1510
https://www.youtube.com/watch?v=dswaElkiRO8	FactCheck.org	Other	8	2	205
https://www.nytimes.com/2021/04/26/us/florida-centner-academy-vaccine.html	Détecteur de rumeurs	Vaccine	7	1	922
https://mcusercontent.com/92561d6dedb66a43fe9a6548f/files/bead7203-0798-4ac8-abe2-076208015556/Public_health_emergency_of_international_concert_Geert_Vanden_Bossche.01.pdf	Détecteur de rumeurs	Vaccine	6	4	487
https://tercalivre.com.br/estudo-frances-aponta-eficacia-da-ivermectina/	Estadão Verifica	Cure	5	5	36
https://www.the-scientist.com/news-opinion/lab-made-coronavirus-triggers-debate-34502	LeadStories	Conspiracy Theory	5	4	2104
https://www.cdc.gov/mmwr/volumes/69/wr/mm6936a5.htm	Détecteur de rumeurs	Other	5	3	1667

Fact-check URL	Topic	Current Week	Previous Week	Total
https://www.factcheck.org/2021/07/scicheck-viral-claim-gets-bidens-covid-19-travel-and-immigration-policies-wrong/	Authorities	106	34	140
https://www.factcheck.org/2021/07/scicheck-viral-posts-misrepresent-cdc-announcement-on-covid-19-pcr-test/	Conspiracy Theory	93	588	681
https://www.factcheck.org/2021/07/scicheck-covid-19-surges-among-unvaccinated-in-florida-contrary-to-baseless-claims/	Spread	80	55	135
https://www.factcheck.org/2021/04/scicheck-idaho-doctor-makes-baseless-claims-about-safety-of-covid-19-vaccines/	Vaccine	63	9	211
https://www.factcheck.org/2021/08/scicheck-pfizer-ceo-got-the-covid-19-vaccine/	Vaccine	59	0	59
https://www.politifact.com/factchecks/2021/jul/23/tiktok-posts/biden-harris-doubted-trump-covid-19-vaccines-not-v/	Vaccine	58	45	124
https://www.factcheck.org/2021/08/scicheck-sequencing-used-to-identify-delta-other-coronavirus-variants/	Other	45	0	45
https://www.factcheck.org/2021/02/biden-hasnt-reduced-covid-19-testing-at-the-border/	Authorities	38	40	307
https://www.politifact.com/factchecks/2021/jul/30/facebook-posts/uk-health-official-misspoke-when-he-said-60-hospit/	Vaccine	34	17	51
https://factuel.afp.com/non-il-ny-pas-800-lits-de-reanimation-en-moins-en-ile-de-france-depuis-mars-2020	Authorities	29	0	95

Figure 4: Number of shares of specific COVID-19 misinformation, fact-checks, and topics, and how this number changed since the previous week.

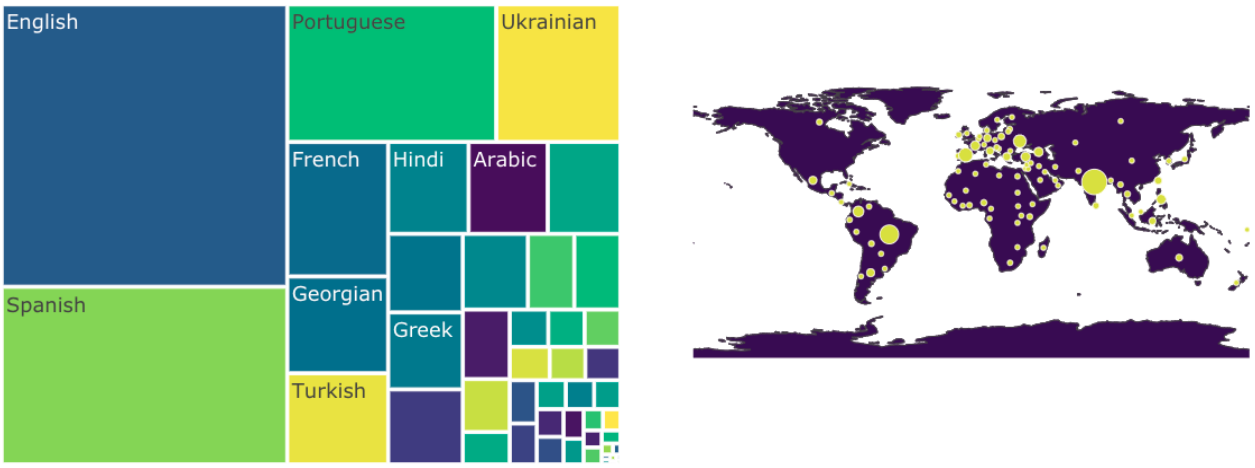

Figure 5:. Relative quantity of misinforming claims written in different languages.

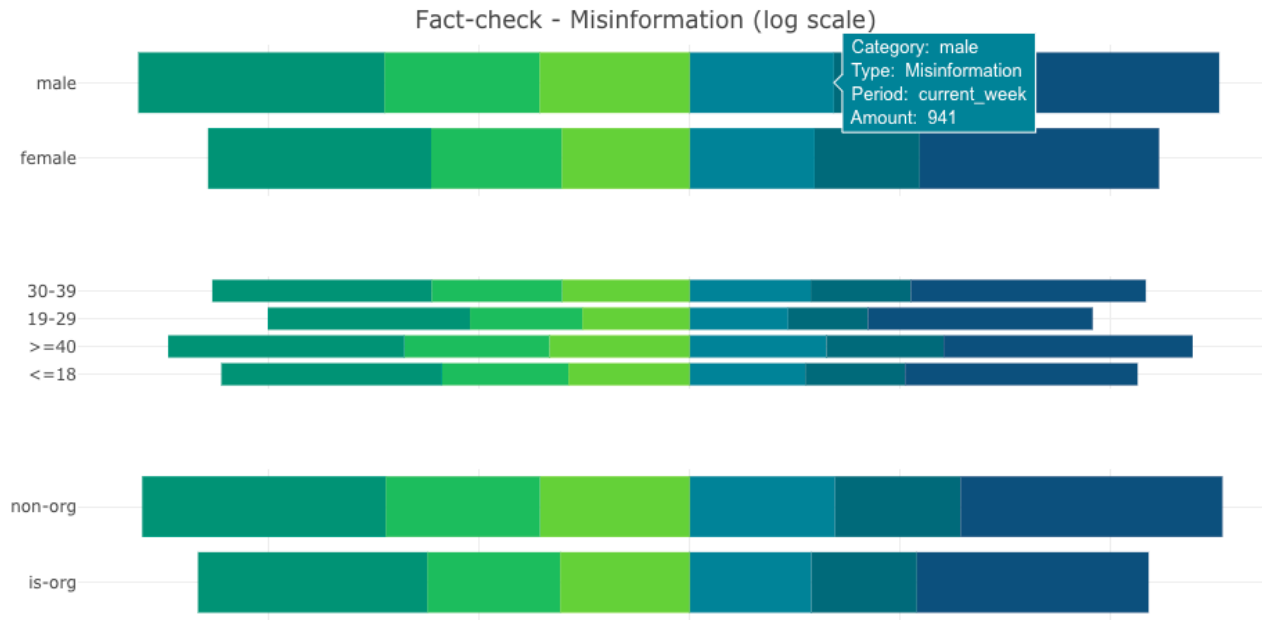

Figure 6: Misinformation and Fact-checking spread across different demographics. Top: Gender, Center: Age group, Bottom: Account type.

Design Update

Following an internal user evaluation, and feedback received after several presentations, including a demonstration to FactCheckNI,[10] the design of FCO was completely revised to achieve a much easier and better user experience. The new version of FCO is to be released later this year, following some further work related to data connectivity. Figure YY shows the new design for the landing page including a new logo.

[10] Fact-Check NI, https://factcheckni.org/

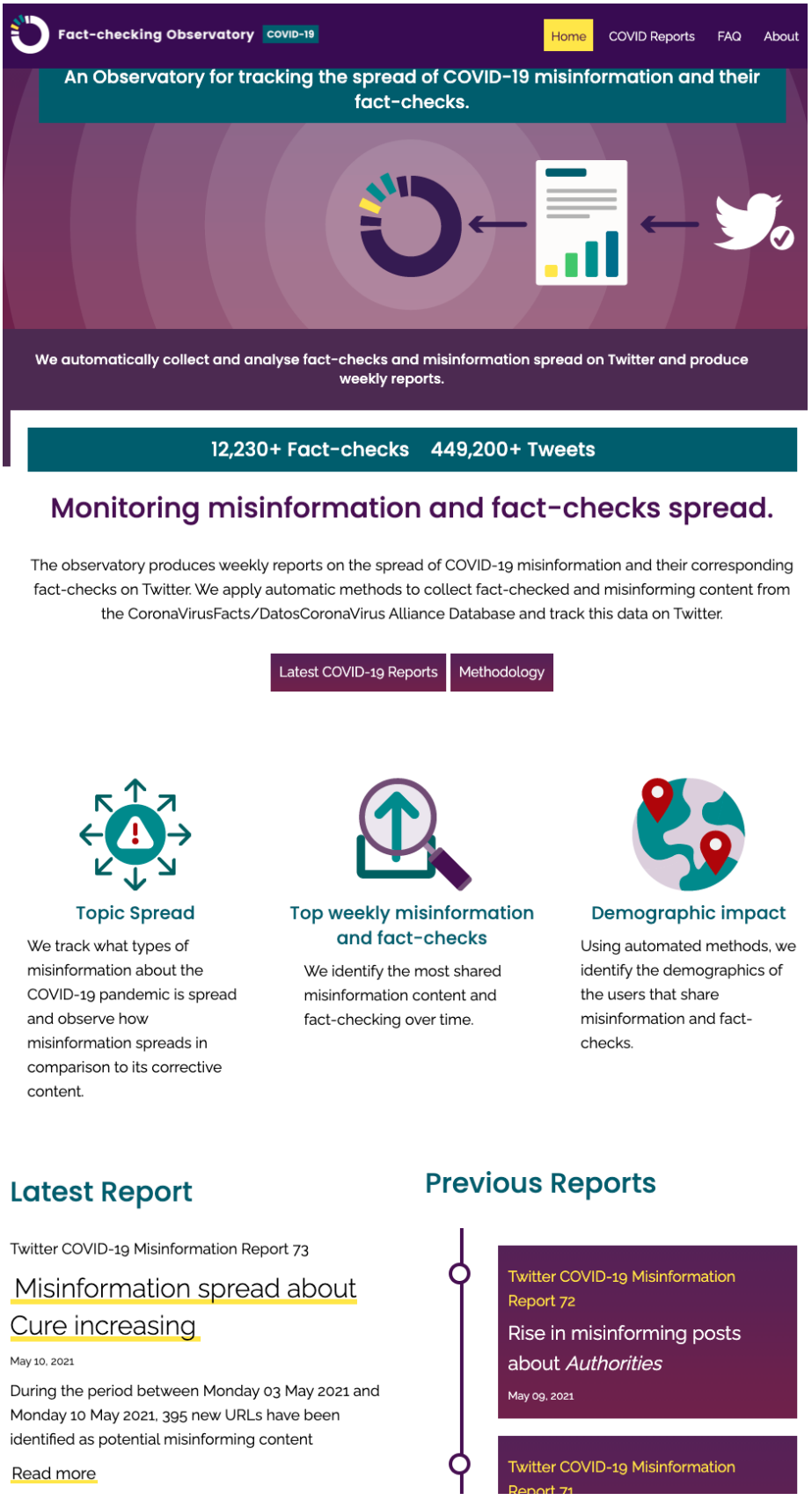

Figure 7: Fact-Checking Observatory new design.

The new design implementation will also include updated reports with additional statistics and graphs for better understanding how misinformation and fact-checks spread over time.

Web mapping application

The FCO's tracking data on COVID19 misinformation and corresponding fact-checks are presented also in a form of a web mapping application enabling (a) geospatial comparison of intensity of misinformation spread and fact-checking over time, (b) analysis of evolution of the most popular misinformation and fact-check topics in various countries (experience.arcgis.com/experience/7207fae129d2440d8cf0c86257172582).

For each month starting from December 2019, the maps present the total number of new occurrences of misinformation and fact-checks, as well as the corresponding most recent top topic. Given that since the geolocalised posts represent only circa 27% of the full dataset, the numbers of occurrences presented on the visualizations are accordingly understated. However, the proportional increase or decrease in numbers, as well as proportion between different countries are respected.

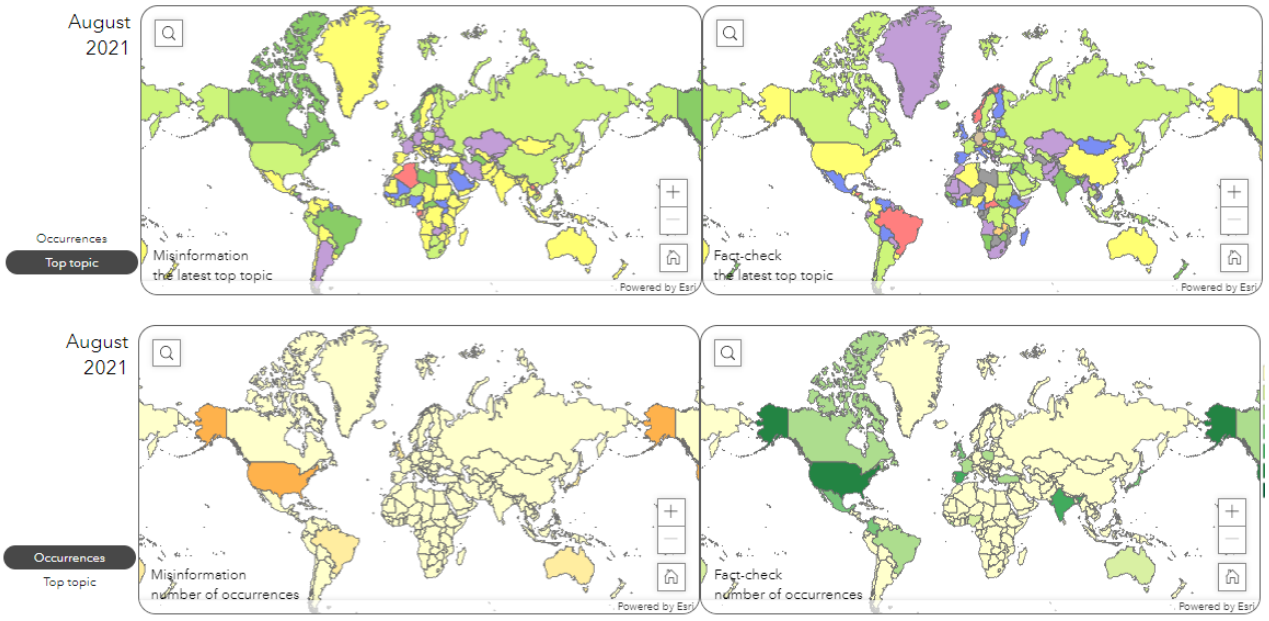

Figure 8:. Interactive maps of misinformation (left) and fact-checks occurrences (right). Top: number of occurrences, Bottom: the most recent top topic.

The set of maps depicting the data for the last completed month are interactive; this allows the user to zoom, navigate and search for a given country.

The new version of the web application is planned for the late autumn this year. The modifications will include i.a. adjustment of temporal resolution of maps to the frequency of the reports generated by the FCO.

3 Classifying COVID-19 posts on social media

Before we introduce our experiments in section 4, it is worth reviewing the ways in which the research community has tackled classification tasks for COVID-19 content on social media. We can divide this research into a few related categories: classification of disease characteristics including symptoms or severity (used to understand the spread of COVID-19) as self-reported on social media, classification of other topics related to COVID-19 (to understand salient issues), user characteristics (to understand more about the public and their concerns around COVID-19 or government guidance) and user sentiment (to understand more about the way the public is feeling about COVID-19 and surrounding issues). In the following subsections, we briefly introduce some of this research and the methods that have been used in the process of classification. We focus on the main categories relevant for our research, topic classification and user characteristics.

Topic Extraction and Classification

A significant body of research on classification tasks for COVID-19 involves extracting topics around COVID-19 from social media posts and tracking these across time. This research is used to understand frequency and temporal characteristics of engagement on the subject of COVID-19. Many approaches use a combination of Natural Language Processing and (semi) supervised approaches, involving the enhancement or scaling of human annotated datasets. For example, Karisani and Karisani (2020) used natural language processing and machine learning techniques to identify posts about COVID-19 infection on the Twitter platform. They first mined Twitter for mentions of COVID-19 using a set of keywords, and then trained a classifier on a subset of those Tweets, using a human annotator. They used seven methods including Naive Bayes, Logistic regression, the fasttex neural network model and four additional models based on the state-of-the-art model Bidirectional Encoder Representations from Transformers (BERT). The authors found that the pre-trained models based on BERT performed best for a binary classification task of positive reports of COVID-19. The paper resulted in the generation of a useful dataset, which the authors have shared on GitHub.[11] There are over 100,000 repository results for COVID-19 on GitHub as of September 2021.

Similarly, hashtag analyses can highlight emerging issues during different phases of the pandemic. In an analysis of more 8.89 million Tweets from January to February 2020, researchers identified more than 2000 hashtags occurring more than 100 times in Tweets related to the coronavirus outbreak (Aguilar-Gallegos et al., 2020). Again, the authors made their datasets available online.[12]

Researchers have been able to build on a large body of open data on COVID-19. One of the goals in all classification tasks is to classify more data. In one study, the authors mined a dataset of 424 million tweets about COVID-19 to "identify discourse around potential treatments." The authors used some classical NLP methods, including the use of lexicons (the authors use drug dictionaries to identify the mention of certain potential treatments), in addition to word-embeddings, spelling-enhancement libraries and keyboard

[11] https://github.com/nkarisan/Covid19_Research
[12] https://data.mendeley.com/datasets/7ph4nx8hnc/1

distance analysis (to enhance with potential misspellings). The use of multiple layered methods resulted in 15% additional data (Tekumalla & Banda, 2020).

Topic extraction is not only about surfacing salient issues. The same methods can be used to analyse word-choices, which can be connected to different perspectives on a topic. Lyu et al. combined demographic data, geo-location data and various ways of referencing COVID-19 on Twitter to track the use of controversial names for the virus. The authors found significant differences in choice of virus names, depending on age, gender, political following, geolocation and other user-level features (Lyu et al., 2020).

User characteristics and COVID-19

The addition of user characteristics, including their interests, other types of information they share and the role they have in public discourse (for example), has provided another layer of understanding to existing approaches for topic analysis. For example, in their comparative analysis of misinformation dynamics on Twitter, Instagram, YouTube, Reddit and Gab during COVID-19, Cinelli et al. (2020) examined more than 8 million COVID-19 comments and posts. The authors found that users who are active on more mainstream social media sites were less susceptible to misinformation. In another study examining discourse about face masks and contact-tracing apps in Switzerland, authors tracked topics in COVID-19 Tweets by three groups of actors: parties, politicians, and what the authors refer to as the "attentive public," (Twitter users following multiple news media). The authors found that different groups were "leading" online discourse about specific issues. For example, discourse around face masks was led by politicians and the public, followed by parties and newspapers (Gilardi et al., 2021).

Researchers also attempt to gather demographic information about social media users, to classify groups of users and to contextualise studies with geographic location, gender or age, for example. For Twitter analysis, many researchers use the geolocation tags provided by users to determine user location. This makes it possible, for example, to identify topics that are more salient in some regions than others. However, it should be noted that Twitter users who geotag their profiles may not be representative of the entire Twitter population (Karami et al., 2021). Researchers have also used analysis of names to determine age (Oktay et al., 2014), gender (Vashisth & Meehan, 2020), ethnicity (Brandt et al., 2020), with varying degrees of success. What's important to consider is that all demographic data that is gathered in this way will be incomplete, but with enough data, may be able to tell a coherent story.

One of the other common ways in which researchers attempt to understand and classify users is through their emotions. While we do not work heavily with this type of analysis, we recognise that it can be helpful for gathering broad-strokes information about public sentiment. Understanding polarity can provide clues toward understanding how citizens understand and appreciate governmental policies and guidelines around COVID-19, the level of anxiety and stress that they feel and even whether or not they are likely to follow advice around prevention of COVID-19 and vaccines. COVIDSenti, for example, is a benchmark dataset for COVID-19-related sentiment, particularly in the early months of the pandemic (Naseem et al., 2021). The developers of this dataset were able to follow lock-down fatigue sentiment, for example, as the first months of the pandemic wore on.

Applying other models on top of this type of analysis can improve its value, particularly when using deep-learning methods. For example, Raamkumar et al. (2020) used the Health Belief Model to classify Facebook comments on official posts by the Ministry of Health of Singapore (MOH), the Centers for Disease Control and Prevention (CDC) in the USA, and Public Health England. The authors found that they were able to accurately classify the four main constructs of the Health Belief Model (perceived severity, perceived susceptibility, perceived barriers, and perceived benefits) with a high level of accuracy. Such studies may be used to understand public concerns and calibrate information campaigns that align with those.

In the following section, we describe our own methodological pipeline for COVID-19 (Mis)information topic classification.

4 Covid-19 (Mis)information Topic Classification

The identification of the different topics related to the COVID-19 pandemic on social media can improve our understanding concerning how specific types of covid-related misinformation spread as well as how general information about the virus is portrayed on social media. In this section, we present ongoing work in automatically identifying multiple topics related to the covid pandemic in Twitter messages using an indirect labelling approach where topic annotations are obtained based on the manual annotation of fact-checking URLs.

In the following sections, we discuss the approach used for creating the training dataset used for creating the proposed topic classifier. Following the presentation of our results, we discuss in more detail different approaches that can be considered in future for approving the topic classification of the proposed models.

Proxied Twitter Covid-19 Topic Classification

The automatic annotation of covid-related topics on Twitter requires multiple steps for creating the necessary models that identify relevant topics. First, we need to collect posts from social media that are globally related to the COVID-19 pandemic and contains multiple kinds of covid-related sub-topics and present both accurate information and misinformation since one important aspect of the HERoS project is to deal with covid-related misinformation. Then, the second step is to both identify the types of topics that are relevant to the COVID-19 pandemic. The third step is to annotate a large amount of the documents collected with the identified topics so that an automatic classifier can be trained. The last step is to clean the annotated data and create different datasets that can be used for training and evaluating the topic classifiers.

The full methodology used for creating the different topic classifiers evaluated in the following section is presented in Figure X. As displayed, our method is based on an indirect labelling method where

fact-checking URLs and their topic labels are used for collecting and indirectly annotating the data used for training the labelling models.

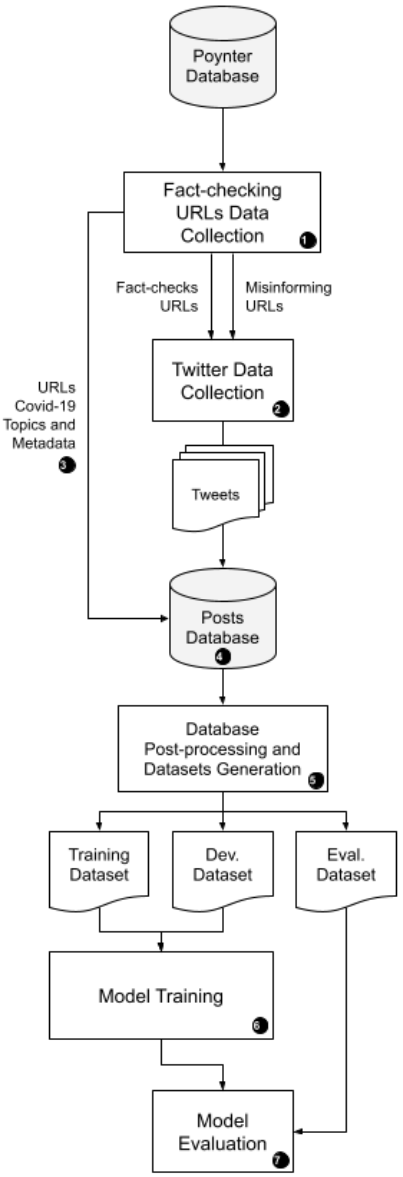

Figure 9: Data collection and processing for generating the dataset used for training the topic classification models.

Dataset and Data Collection

The methodology used for collecting the necessary data uses the same methodology as the approach used for obtaining data for the Fact-checking Observatory. We collect the tweets that mention covid-related misinformation and fact-checking URLs using data from fact-checking organisations. Then, we use the labels associated with such URLs as proxy annotations for the collected tweets. We collect data spanning a period between the 1st of December 2019 until the 17th of August 2021.

Fact-check URLs and topic dataset

Rather than collecting a database of Tweets based on covid-related hashtags, we look for tweets that mention URLs that are known to discuss the COVID-19 pandemic. Since as part of the HERoS project we are also interested in understanding how misinformation spreads, we bootstrap our data collection using URLs from fact-checkers that are related to the COVID-19 pandemic.

The first step for collecting the tweets is to obtain a list of relevant fact-checking and misinforming URLs. The dataset of fact-checks and misinforming URLs that we use comes from the Poynter Institute's International Fact-Checking Network (IFCN). This is the same data that is used for the Fact-checking Observatory. The dataset is extracted from Poynter's COVID-19 specific fact-check alliance database.[13] The Poynter database aggregates fact-checking reviews from more than 100 verified fact-checking websites around the world about issues surrounding COVID-19, and consists of fact-checking URLs, the reviewed URLs, and the fact-checker rating (e.g.,False, True).

Each URL aggregated by the Poynter Institute is associated with a rating indicating the validity of a particular claim. Besides such information, the URLs obtained through the Poynter Institute are also linked to a covid-relateed topic. These topics can be used for indirectly annotating the twitter posts collected that mention the URLs found in the Poynter database.

For the purpose of the COVID-19 specific fact-check alliance database, the following 9 topics are identified in relation to the pandemic:[14]

- Authorities: Information relating to government or authorities communication and general involvement during the COVID-19 pandemic (e.g., crime, government, aid, lockdown).
- Causes: Information about the virus causes and outbreaks (e.g.,China, animals).
- Conspiracy theories: COVID-19-related conspiracy theories (e.g., 5G, biological weapon).
- Cures: Information about potential virus cures (e.g., vaccines, hydroxychloroquine, bleach).
- Masks: Information relating to the usage of masks for protection against COVID-19.
- Vaccine: Information about the development, effectiveness and usage of vaccines against COVID-19.
- Spread: Information relating to the spread of COVID-19 (e.g., travel, animals).
- Symptoms: Information relating to symptoms and symptomatic treatments of COVID-19 (e.g., cough, sore throat).
- Other: Any topic that does not fit the aforementioned categories directly.

[13] Coronavirus Facts Alliance, https://www.poynter.org/coronavirusfactsalliance.
[14] The IFCN database does not give any explicit description of each category. The description of each category is derived based on the claims in each topic.

Twitter dataset

Using the URLs collected from the Poynter institute and the associated topics, we create the annotated Twitter dataset by searching the URLs occurrences on Twitter a posteriori using a crawler based on the TWINT Intelligence Tool.[15]

Using this method, we collect a total of 481,066 posts covering the aforementioned 9 topics. The distribution of posts for each topic is as follows: (1) Authorities: 128,170; (2) Causes: 34,601; (3) Conspiracy Theory: 57,576; (4) Cure: 86,038; (5) Masks: 11; (6) Other: 108,263; (7) Spread: 34,200; (8) Symptoms: 2,925, and; (9) Vaccine: 29,282.

Topic Classification Models

Using the database of annotated tweets, we can now generate the datasets used for training and evaluating different models used for classifying the social media posts. Before selecting a set of candidate classification models, we need to preprocess each post to avoid overfitting. We also need to take great care concerning how the training, evaluation and test datasets are created so that the tweets concerning the same URLs are not shared across each dataset.

Database Post-processing and Dataset generation

Before creating the training, development and evaluation datasets, we perform some post-processing on the collected data in order to reduce overfitting of the classification models. For each tweet, we replace the hashtags, mentions and URLs with placeholders so that the trained models are not biased towards specific user mentions and hashtags. The URLs are removed so that the models do not focus on the URLs for identifying the COVID-19 topics.

Besides the previous preprocessing, we also remove any potential duplicates such as retweets and only keep posts that contain at least five words.

Since the distribution of topics for the retrieved posts is not uniform, we decide to only focus on the topics that appear sufficiently for training the models efficiently. For a topic to be kept, it needs to appear in at least 5% of the collected data. Following this approach the following 7 topics are kept: (1) Authorities; (2) Causes; (3) Conspiracy Theory; (4) Cure; (5) Other, and; (6) Spread, and; (7) Symptoms.

For training the different models, we need to generate a training, testing and evaluation dataset. In order to avoid any leak between each dataset, we make sure that the same misinforming and fact-checking URLs do not appear across each dataset.

[15] TWINT Intelligence Tool, https://github.com/twintproject/twint.

After filtering the collected posts, we obtained 280,270 posts. The distribution of posts for each topic is as follows: (1) Authorities: 84,165; (2) Causes: 20,613; (3) Conspiracy Theory: 28,545; (4) Cure: 52,063; (5) Other: 60,756; (6) Spread: 17,857, and; (7) Vaccine: 16,271.

We try to generate the train, development and test datasets so that they respectively represent around 80%, 10% and 10% of the original dataset. We try to keep similar topic proportions across each dataset. However, since we have to make sure that URLs are not shared between the subsets, we obtain some slight variation in size and topic across the dataset despite using an iterative topic allocation approach while generating each dataset. Finally, it is important to note that we do not balance the dataset as we use weighted metrics for training and evaluating the trained models. The proportion of the different topics for each dataset is displayed in Figure X.

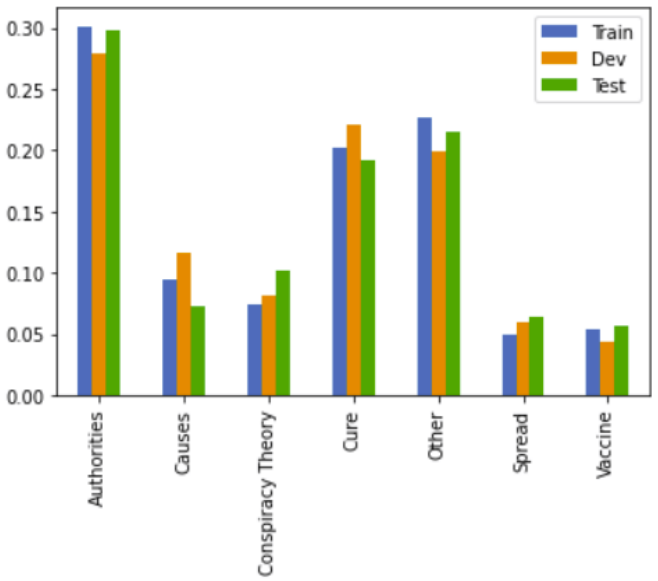

Figure 10: Proportion of posts with a given topic for the final train, development and test datasets.

Model Training and Evaluation

We train multiple classical classifiers and Deep Neural Network (DNN) models. For the classical model we train the following multiclass models: (1) a random forest model; (2) SVM; (3) naive bayes, and; (4) a logistic regression model. For the DNN models we fine tune the following models: (1) BERT (uncased); (2) Distil-BERT (uncased); (3) RoBERTa; (4) XLM RoBERTa, and (5) Glove.

For the classical models, we use both unigrams and bigrams and use TF-IDF normalisation. For the DNN, we train the models on 5 epochs using ADAM optimisation.

For each model, we report the micro, macro and weighted F1, Precision (P) and Recall (R) values. For the best model we also report these scores for each class.

Classification Results

The results of the multiclass classification models are reported in Table 1. The results for all the models show that in general the overall classification of the topics is hard with the best macro average F1 of around 32% for the Distil-BERT model.

	Micro Average			Macro Average			Weighted Average		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Random Forests	-	-	-	0.221	0.232	0.204	0.221	0.232	0.205
SVM	-	-	-	0.288	0.301	0.292	0.288	0.301	0.292
Naive Bayes	-	-	-	0.280	0.280	0.280	0.281	0.281	0.281
Logistic	-	-	-	0.295	0.314	0.314	0.295	0.314	0.301
BERT	0.334	0.334	0.334	0.303	0.334	0.314	0.303	0.334	0.314
Distil-BERT	0.347	0.347	0.347	0.304	0.347	0.318	0.304	0.347	0.318
RoBERTa	0.331	0.331	0.331	0.301	0.331	0.313	0.301	0.331	0.331
XLM RoBERTa	0.319	0.319	0.319	0.288	0.319	0.298	0.288	0.319	0.319
Glove (untuned)	0.160	0.160	0.160	0.145	0.160	0.141	0.145	0.160	0.145

Table 1: Classification evaluation results for different models.

Overall, it seems that DNNs perform more accurately than more classical models with Distil-BERT providing the highest macro average F1. The Glove model appears to be the worst model since it is not fine-tuned and depends on pre-existing embeddings that may be not well suited for Twitter data. Logistic regression appears to perform quite well compared to the best model with a macro average F1 of around 31%. This result seems to indicate that more classical methods may be sufficient when needing less computationally intensive results.

The detailed results for the Distil-BERT model displayed in Table 2 show that some classes are easier to predict compared to others. In particular, it seems that the *causes* topic is particularly hard to predict whereas the *Vaccine* topic is much easier to predict. This result may be explained by the fact that viruses causes may be discussed across different covid-related topics whereas vaccines-related topics are much more obvious to classify with very clear terms used.

Class	Precision	Recall	F1-score
Other	0.3204	0.3907	0.3521
Authorities	0.3873	0.5462	0.4532
Vaccine	0.4738	0.5778	0.5206
Conspiracy Theory	0.1626	0.1518	0.1570
Spread	0.2836	0.1998	0.2345
Cure	0.4320	0.5440	0.4815
Causes	0.0680	0.0173	0.0276

Table 2: Per class classification results for the Distil-BERT model.

The confusion matrix displayed in Figure 11 shows the proportion of predicted labels for the Distil-BERT classifier against the true labels (the labels obtained from Poynter). This allows us to see how the algorithm is performing, in particular, if it is confusing classes in the predictions. From the confusion matrix, we can see which topics were more successfully classified (posts about Authorities, Cures and Vaccines) and which were less successful (Causes and Spread).

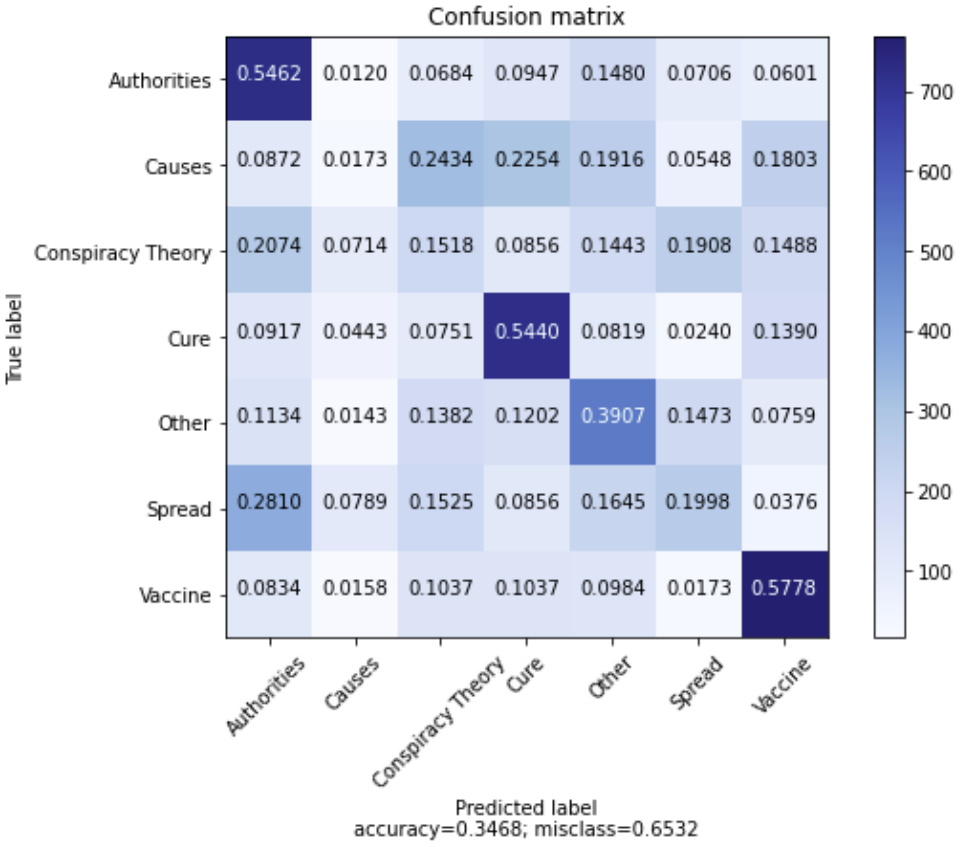

Figure 11: Confusion matrix for misclassified posts

Looking at the confusion matrix we can observe what topics are the most likely to be confused with others by the classifier. In particular, we can observe that posts about *authorities* tend to be easily confused with posts about *conspiracy theories* and *vaccine spread. This* observation may be simply due to how related these topics are in practice with conspiracy theories, authorities and virus spread generally linked together. In general these results show that for improving the proposed classifier, we need to better understand the relation between the topics as well as the common mistakes made by the classifier for each topic.

Qualitative Error Analysis

The closeness of the different topics appears to be key in understanding why the proposed classifiers fail. In order to improve the proposed classifiers including our best model, we need to better understand why and where misclassifications occur. In this section we perform some qualitative analysis by analysing a sample of 140 posts that were misclassified (20 from each class).

In the following section we describe some of the qualitative explanations that we were able to derive for some of these misclassifications. Note that we remove the category of “other” for the moment, as these errors have multiple general explanations. In addition, we have removed all references to @mentions, numbers, and URLs, denoted by the use of the ampersand symbol and the category of item that was removed. We leave them in the examples, so that the reader can see how much information is missing from the misclassified posts.

General explanations

Some errors can be found throughout the sample, across various topics. We identified the following general errors:

1. Incorrect labelling (general) - Sometimes Poynter labels are inaccurate and our labels are more closely aligned with the post. In order to improve on such generic cases, more disparate labels may be necessary or allow the model to return multiple labels for a given post.

For example, we classified the following post as “cures” and poynter as “causes”:

> *"$hashtag$ este contenido ya no aparece en las redes sociales del $mention$ la $mention$ descarta la efectividad de cualquier remedio casero para combatir el covid19 revisa url url"*

In this example, the post is a debunking post, in which the user is warning their audience about an apparent post from the army of Ecuador recommending chewing ginger to prevent virus replication in the throat. The user mentions that the World Health Organisation has “ruled out the effectiveness of any home remedy to combat COVID-19” and points to a URL with many additional debunks and fact-checks about COVID-19. As Poynter labels arise from human fact-checkers labelling URLs, this error is to be expected.

To provide another example, we labelled the following post as “authorities” and the Poynter label was “causes”:

> *"$hashtag$ $mention$ de sao paulo asegur que el secretario de comunicacin de la presidencia de brasil $hashtag$ dio positivo para $hashtag$ el presidente $mention$ dio negativo para $hashtag$ url url"*

This post is discussing which authority figures in Brazil have tested positive for the virus. It also refers to the communication of a municipality, Sao Paulo. In our view, such a post is about authorities, more than it is about causes. Upon examination of the URL, the page appears to be debunking several different rumors

about COVID that are circulating in Ecuador, including misinformation about causes and cures. This may provide a partial explanation for the Poynter label.

2. Post-URL confusion- Since we must remove the hashtags from the analysis to avoid bias and we also must remove the URLs from our analysis and treat them as unknown, we are not able to see if the URLs address the same type of information provided in the post. Nor are we able to see the full informational content of the post (with hashtags). So, occasionally, this leads our classifier to misclassify posts.
For example, we annotate the following post as “authorities” and Poynter as “spread”:

> *great to see proper facts checking excellent work $mention$ debunked is $number$ really a normal year for deaths from respiratory illnesses url*

From the post alone, we have reference to fact-checkers, which may be the reason why the algorithm classed it in this way. With the URL, we can see that the fact-checked content has to do with the impact of COVID-19 on deaths (inadvertently spread and severity, although the Poynter label is also not particularly clear in this case).

To provide another example, we labelled the following post as “conspiracy” and Poynter as “causes”:

> *$mention$ i see more lies and bs coming from you alleged fact checkers than from mainstream news for example this page absolute bs who are you funded by bill gates url*

Presumably, in the URL, the topic is related to “causes”, but our algorithm picks up on “funded by bill gates” as linking the post to conspiracy theories. This could also be an example of multiclass belonging, as described below in number 6. In practice, it is important to note that the model would also match any URL already observed in the Poynter data. As a result, for an already known URL, the classifier would always allocate the correct label through direct URL matching.

3. Slang/Symbols/Underrepresented languages - Languages that combine several languages, or that are heavily related to other more highly represented languages (such as Catalan, Austrian German and Danish, for example) appear more difficult to classify. In addition, languages that use heavy slang or incorporate additional signs and symbols within the text make it difficult to classify text.

4. Vague - Vague or incomplete sentences that let the additional content (URL or image) "speak for itself" were difficult for us to classify, particularly because we were unable to make use of the URL in the classification problem. This problem arises quite a bit in many classification problems on Twitter. In this context using an additional model that classifies known hashtags or is able to analyse images may be beneficial. For example, we misclassified the following posts:

> *"$mention$ $mention$ $mention$ $mention$ $mention$ $mention$ $mention$ $mention$ incorrect url"*

> *"Polifact also back legalin and APnews on the subject."*

5. Multiples of the same misclassification/disputable classification - when the same story appears more than once in our database and is misclassified multiple times. For example, the following post appears more than 60 times in our sample (we coded this post as “spread” and Poynter as “other”):

> *“hmm seiu union in california suddenly finds mysterious stash of $number$ million face masks $number$ days after ag bill barr announces theyre going after hoarders url $mention$”*

6. Multiclass Belonging - Posts with multiple sentences create difficulty for classification, because they may include more than one topic in the full statement. Allowing our classifier to generate more than one label at a time mwy improve our results. For example, we labelled the following statement as “cure” and Poynter as “causes”:

> *"$hashtag$ no difundas cadenas engaosas no se ha comprobado la efectividad de remedios caseros para evitar el contagio de $hashtag$ la $mention$ insiste en que la higiene de las manos es un mtodo para evitar la propagacin $hashtag$ y $hashtag$ url url"*

This statement encourages users not to spread misinformation about the effectiveness of home remedies to prevent the spread of COVID. The post points to a URL in which hand hygiene is suggested as a preventative method.

Topic Specific Explanations

Authorities

In our confusion matrix, we can see that the label “authorities” is most often confused with “causes”. However, we observed other misclassification patterns related to what the authorities are saying, or what is being said about them. Confusion between classes of “authorities” and other classes about spread, cures, vaccines, etc. has to do with the fact that authorities speak about those issues and the public responds to those statements. This can make it difficult to analyse the topic: For example, we annotated the following example as “spread” and Poynter as “authorities”.

> *"marco rubio says anthony fauci lied about masks fauci didnt the message on masks was primarily about preserving a limited supply for health care workers who were at especially high risk of exposure"*

Similar to the above, when a piece of misinformation is about an Authority, it can confuse the labelling process. In the following case, we have an authority that has been the subject of misinformation about the vaccine (we coded as “vaccine”, Poynter as “authorities”):

> *"president cyril ramaphosa did not receive his covid 19 vaccine with a capped needle url url"*

However, this error may also be due to the fact that the “vaccine” label was introduced only more recently by Poynter. It is possible that the vaccine option was not available when the URL was initially annotated. Confusion in annotation schemes, especially those that are evolving can account for some errors.

Causes

Causes were confused mostly with spread in our analysis, but it appears to be the label that we have the most difficulty with. That could be because it looks like so many other potential labels. For example, this one was labelled by our classifier as “spread”, and Poynter as “causes”:

> *"although us is pushing now for a leak out of a wuhan lab the recent research also says that a virus found in pangolins banned smuggled in wuhan markets is a very close match to covid 19 not clear but could be mutated once in humans may not reoccur url"*

On one hand, this post could be interpreted as thinking about where the virus came from (wet markets). On the other hand, it can be viewed as a statement about how COVID-19 and similar viruses spread to human populations. Confusion also happens between causes, spread and conspiracy theories, particularly around where COVID-19 can be found (which can lead to spread). This highlights again the issue of having topics

that are highly related. For example, the following post was labelled by our algorithm as “other”, whereas Poynter labelled it as “causes”:

> *"a newspaper clipping claiming that the bihar health department has found $hashtag$ in poultry chicken samples it tested is fake $mention$ $hashtag$ $hashtag$ $hashtag$ $hashtag$ $hashtag$ $hashtag$ $mention$ url"*

Conspiracy theory

Conspiracy theories are most often confused with causes and cures in our matrix. This makes sense, because the public has the most insecurity and is lacking the most information around where this virus came from, how to prevent it and how to fix it. In addition, the definition of conspiracy theory may not be immediately clear. There are many alternative health positions (in particular around COVID-19 as a bacteria vs. a virus) that one could classify as conspiracy theories because they try to call into question vaccination programs and link to theories of why governments may want to vaccinate their citizens (to inject other substances into the bodies of citizens for example, or to support “Big Pharma”). Causes can also be confused with conspiracies, for example, when they relate to the origins of COVID-19 (around which there are many theories). The following posts were labelled by our algorithm as “conspiracy theory” and by Poynter as “causes”:

> *“the search for the origin of $hashtag$ continues pangolins human transmission prime suspect but nothing definitively proved yet url”*

> *“closest match to the human coronavirus has been found in a bat in chinas yunnan province study published feb $number$ rd found that the bat coronavirus shared $number$ of its genetic material with covid 19 coronavirus bats could have passed the virus to humans url”*

In many of the English-language posts that were misclassified, either the bat or the pangolin-transmission theory appears. Perhaps Poynter is less inclined to label such posts as conspiracy theories, despite their connection to conspiracy theories about the origins of the virus. However, it is also worth noting that a large number of the misclassified posts that we labelled as “conspiracy theory” are in languages other than English.

Spread

In our matrix, the label spread was more often confused with the labels of vaccine, cures and causes. Spread overlaps to some degree with origin stories of where COVID-19 came from and why it spread (and conspiracy theories about this), as described above. Confusion with vaccines may also arise through multiclass belonging where the post discusses the impact of vaccines on the spread of COVID-19. For example, our algorithm labeled the following post as “spread” and Poynter as “vaccine”-related:

> *“$mention$ $mention$ $mention$ $mention$ your figure is incorrect however death is not the only problem hospitalization for long periods is a serious issue and strains heath services and recovery doesnt mean return to full health url”*

The post is in reference to a post by other users about the efficacy of vaccines. Since vaccines' goal is to reduce the spread of the virus, it may also explain the confusion between the two labels.

However, we identified some errors that do not appear to have a solid explanation. For example, we classified the following posts as “spread” and Poynter as “vaccine”:

> *“when the history of this madness is written reputations will be slaughtered and there will be blood in the gutter url”*
>
> *“they will be ripped to shreds and slaughtered when all this ends url”*
>
> *“$mention$ my new hero actually i can t remember who my previous hero was lol url”*

The URLS all link to an article by Daily Expose on Dr. Roger Hodkinson and statements he apparently made on how vaccines cause male infertility and could harm pregnancies, among other positions against the provision of vaccines. This same URL is linked to several misclassified posts.

Cures

In the confusion matrix, the label of “cures” was confused mostly with “spread” and “causes”. We have some of our better results for this class, potentially because there are specific terms and phrases associated with home remedies and potential pharmaceutical treatments. However, there is still much potential for misclassification. We've already mentioned how cures can be confused with conspiracy theories about spread and causes. For our “cure” label, there are many posts about the transmission of the virus that are more accurately classified as “causes” or “conspiracy theories”. For example the following post was classified as “cure” by our algorithm and “causes” by Poynter:

> *$number$ the well known hazards of coronavirus vaccines falsely claims that flu vaccination increases the risk of coronavirus infection url your ads are funding this $hashtag$ url*

Some misclassifications have to do with mentions of ivermectin and hydroxychloroquine acid as part of conspiracy theories (where people believing in such cures then think that the government is trying to hide key characteristics of COVID-19 or overstate its severity). Others result from some alternative health positions around whether or not COVID-19 is caused by bacterial infection vs. a viral infection. For example, our algorithm labelled the following post as “cures”, whereas Poynter labelled it as “causes”:

> *“autopsies performed by italian pathologists supposedly uncovered that covid 19 is not pneumonia but it is disseminated intravascular coagulation thrombosis which ought to be fought with antibiotics antivirals anti inflammatories and anticoagulants url”*

We also see users posting frustrated messages as they exchange scientific evidence around cures. This evidence can cover a lot of different subjects and some of the ways in which it is presented could be confusing for a classification task. Many misclassifications arise from having little information about what the scientific evidence entails from the post: For example:

> "$mention$ $mention$ $mention$ $mention$ and i ll just leave his right here if only you crazy leftists spent more time actually research the virus as you do for stupid memes url"

Our algorithm labelled this post as having to do with “authorities” (possibly because of the word “leftists” having a political origin), and Poynter coded it as “cures”. As the URL leads to a website investigating use of hydroxychloroquine, it would appear that Poynter's annotation is correct.

Vaccine

In the confusion matrix, the “vaccine” label is confused with causes and spread, as we have already described above. This can be impacted, as we mentioned previously, by the late addition of the “vaccine”

label. However, another way in which confusion arises is when users link other types of vaccines to COVID-19. For example, the following post was coded by our algorithm as being “vaccine” related and by Poynter as “causes”.

> *“false claim $hashtag$ the $hashtag$ vaccine increases your risk of covid truth the $mention$ has concluded that the flu shot will not make you more vulnerable to other respiratory infections $mention$ debunks this myth here 9 x url”*

Confusion may arise additionally when users post about alternative health positions around the causes and spread of the virus, as well as conspiracy theories about government positions on the efficacy of vaccines. For example, this post was also annotated by our algorithm as “vaccine” related, and by Poynter as “causes”:

> "*covid 19 une dcouverte majeure lisez c est la fin le covid est vaincu"*

The post says that COVID-19 is defeated and leads to the following URL:

https://israelmagazine.co.il/covid-19-une-decouverte-majeure/,

This URL is an article about how COVID-19 is caused by a bacterial infection and not by a virus, making the vaccination program obsolete. This, in turn, can be related to conspiracy theories about government positions on the efficacy of vaccines, demonstrating that this could also be an example of the multi-class problem.

Discussion and Future Work

The error analysis shows some insights concerning how the topic classifier could be improved and refined. First, it seems that the COVID-19 topics may be too related to each other, meaning that the topics are hard to distinguish and may behave more like topic hierarchies. For dealing with such an issue, the definition of more disparate topics may be needed. Unfortunately, this approach would need new topics to be defined and manual annotations. Moreover, the small length of Twitter posts may remain an issue for distinguishing the topics even if their relationship is not as close as the current ones.

Rather than sticking to a multiclass scenario where posts are only associated with one topic, considering the relationship between the topics and allowing the classifier to behave similarly to a multilabel classifier where more than one topic can be assigned to a given Twitter post may be beneficial. Although reannotating the Poynter URLs with multiple classes may be a possibility, a relatively straightforward approach would be to make the current model a conformal model (Balasubramanian et al., 2014), where it is possible to predict a set of classes according to a specified confidence interval. Using this approach, it becomes possible to predict a set of topics for each post and therefore increase the effectiveness of the proposed model.

Another approach for improving the classification results would be to drop some of the classes that are too generic or ambiguous such as the *other* class, leading to an overall higher accuracy of the model.

For training the models, we have intentionally left out the hashtags, mentions and URLs for avoiding the model being overfitted. However, such information is useful when available. An approach for integrating back such information into the topic classifier could be through the use of multiple sub-classifiers and a

meta-classifier so that the presence of such indicators can be used when available without overfitting the base model.

For the creation of the classifier, we have also purposely removed the URLs from which we already know to be allocated to particular categories. In practice, the classifier would also match any URLs from our database and classify new URL mentions appropriately as new URLs are collected as part of the FCO.

Other issues and limitations associated with the current approach can be traced with the multilingual nature of the data as well as the way the data is annotated. First, due to the presence of multiple languages, our model performs better in English compared to less present language. This can be improved by using multilingual transformer models or using automatic translation (Khare et al., 2018). Our proxied approach has the issue of only working on fact-checked content and Tweets that contain URLs. This means that our model may also be limited when dealing with content not linking to URLs as they may use a slightly different language that may be unrecognised by our classifier.

As part of the HERoS project, we plan to improve the classification model based on the aforementioned observations. In particular, future work will investigate only keeping specific topics as well as making our model a conformal classifier so that multiple topics can be derived for a given post. Finally, we will investigate creating an API for the classifier so it can be used for identifying topics in new Tweets.

User Descriptions Co-Occurrence Hashtags Analysis

In order to understand the relation between user orientation towards COVID-19 misinformation and information in general, we perform some initial analysis concerning how users define themselves through the use of hashtags on their Twitter profile description and the type of information they discuss in their posts. Using this approach, we aim at better identifying key values and demographics, their intrinsic relation (i.e., co-occurrences) and how they are associated with either sharing misinformation or fact-checks.

Data Collection and Co-Occurrences Generation

As with the topic classification task presented in the previous section, we rely on the data collected as part of the FCO. Using all the posts collected by searching misinforming and fact-checking URLs on Twitter, we obtain a list of users that have either shared misinformation or fact-checks. From this list of users, we then compile a list of identifiers from their self-produced Twitter bios. For each of them, we obtain their Twitter profile description and extract all the hashtags present before generating all co-occurrences for each hashtag.

After obtaining all the co-occurrences for each user profile, we need to identify if user profiles are more likely to share misinformation or fact-checks overall (it is important to note that users that share fact-checks may also share misinformation as part of their discussion). In order to perform this task, we calculate the average inclination towards misinformation or fact-checks using the normalised claim review annotations present in the posts collected as part of the FCO. As a result, we classify users as misinforming users if they have an average normalised claim review score less than or equal to 0, and classify users with an average score above 0 as informing users.

After obtaining the user inclinations, we generate two occurrence graphs based on the hashtags they use to describe themselves. One graph is for misinforming users, and one is for users that share fact-checked content. The graphs represent the connections between hashtags in users' biographies, where the width of the link connecting two nodes (hashtags) indicates how often a connection is observed (as more occurrences are observed, the links become thicker) and the size of a node indicates how many different hashtags are linked to a particular hashtag (in-degree). Our analysis is performed on a total of 265,948 user profiles.

The resulting images are quite large. For this reason, we show a version of the image, annotated by hashtag grouping and then a selection of other images that go more deeply into detail. In Figure 12, we see the hashtag co-occurrences for users sharing more misinformation. We see three primary groupings in this image, which can be associated with the broad categories annotating each group. The first obvious group are individuals who have description hashtags associated with social justice topics, such as feminism and Black Lives Matter. The second group (from top to bottom) are those who have hashtags in their bioline related to technology, such as cryptocurrencies, A.I., and blockchain. The third large group are those with more conservative political hashtags and those related to patriotism, being pro-Trump, defending the US constitution, and religion.

It's not surprising to see political polarisation akin to that in the United States in this network graph. Demographic analysis of Twitter users[16] suggests a younger, more affluent, millennial user-base, with a large presence of users from the United States. What is more interesting, aside from the relative dearth of non-politically coded identifiers, is the clear coalescence of both conservative and liberal groups, as research in this area has been difficult to interpret. Guess et al (2018) found that conservatives and older users were more likely to share news from disreputable sources, but that older users were also more likely to share facts. Some studies show that conservatives share more misinformation (Ecker & Ang, 2019; Grinberg et al., 2019), while other studies have argued that confounding factors, such as perceived bias in the media and in fact-checking organisations (Allcott & Gentzkow, 2017) or information processing tendencies of conservative versus liberal individuals (Harper & Baguley, 2019) better explain why research implicates conservatives more than liberals in sharing misinformation. Harper & Baguley (ibid), for example, argued that liberals and conservatives are vulnerable to misinformation for different reasons. They found that the greater the partisan attachment (on either side), the more willing individuals appear to be in engaging in ``cognitive distortion'' to protect their views. Earlier research implicated dogmatism, religious belief, and delusional ideation with belief in fake news (Bronstein et al., 2018), as well as overconfidence in one's knowledge and a lack of critical analytic skills (Pennycook & Rand, 2018). Perhaps

[16] https://www.omnicoreagency.com/twitter-statistics/

these potential factors exist regardless of partisan attachment, but are expressed in different ways and toward different subjects (as we hope to research in the immediate future).

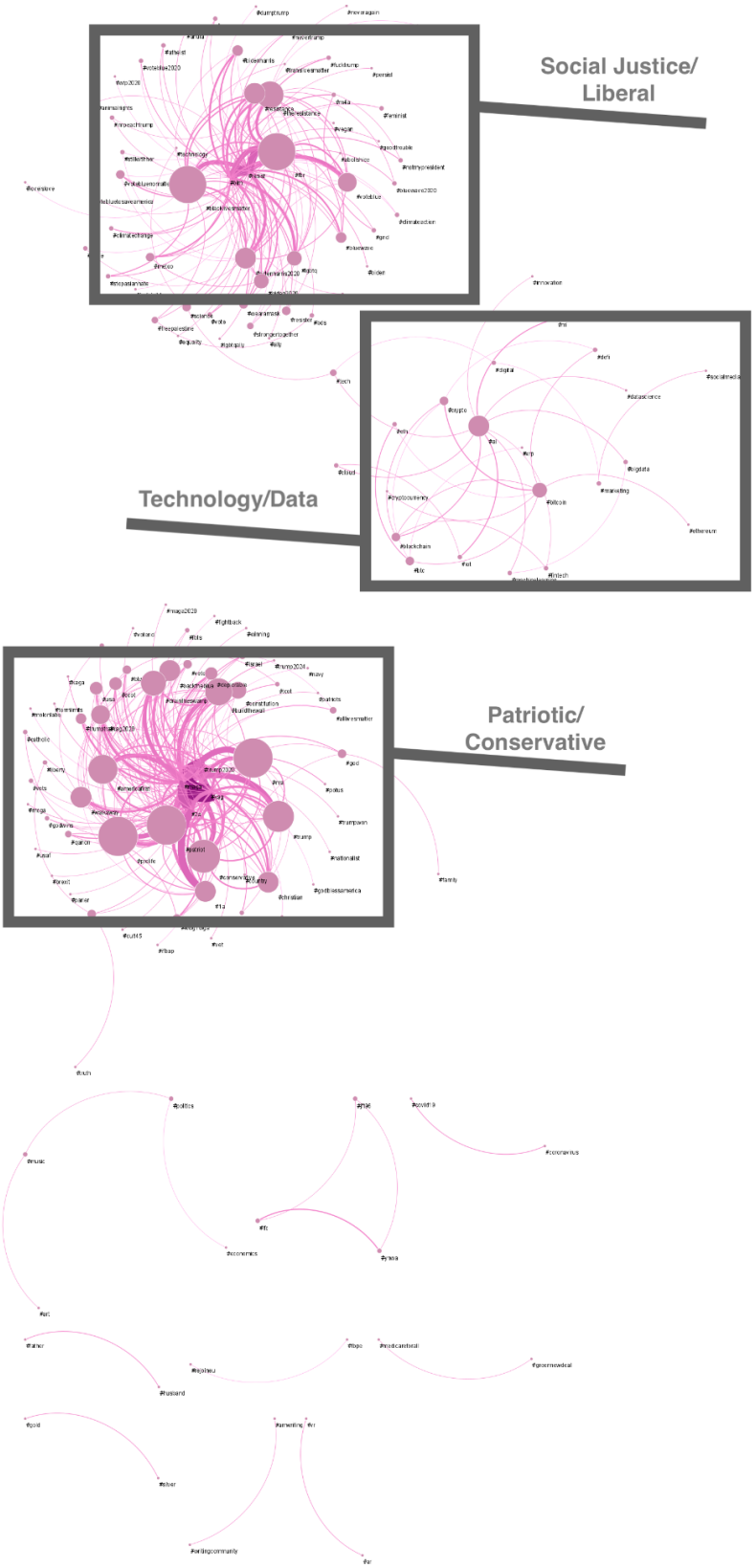

Figure 12: Hashtag Co-occurrences of users sharing misinformation

Fact-checking is a bit more nebulous, in terms of what we know and don't know about fact-check-sharing behaviour on social media. Fact-checking itself is viewed by some individuals and groups as problematic, encouraging partisanship and polarisation, and promoting dichotomic understandings of science and scientific dissent (Clarke, 2021). Earlier studies have found that United States Republicans perceive fact-checking in a more negative light as compared with Democrats (Nyhan & Reifler, 2016; Guess & Nyhan, 2017; Amazeen et al., 2019). Later studies indicated that Republications can support fact-checking, but not about certain topics, like previous president Donald Trump (Rich et al., 2020). Amazeen et al. (2019) found that users share fact-checks, predominantly, due to a need for orientation and that individuals who are liberal-leaning or older were more likely to post a fact-check. They also found evidence of users posting fact-checks for "attitude reinforcement", more than resisting misinforming narratives. In Figure 13, we have the broad picture of users sharing fact-checked URLs. In this graph, we see evidence of the above, in that the liberal group is much more closely implicated in sharing fact-checks on Twitter.

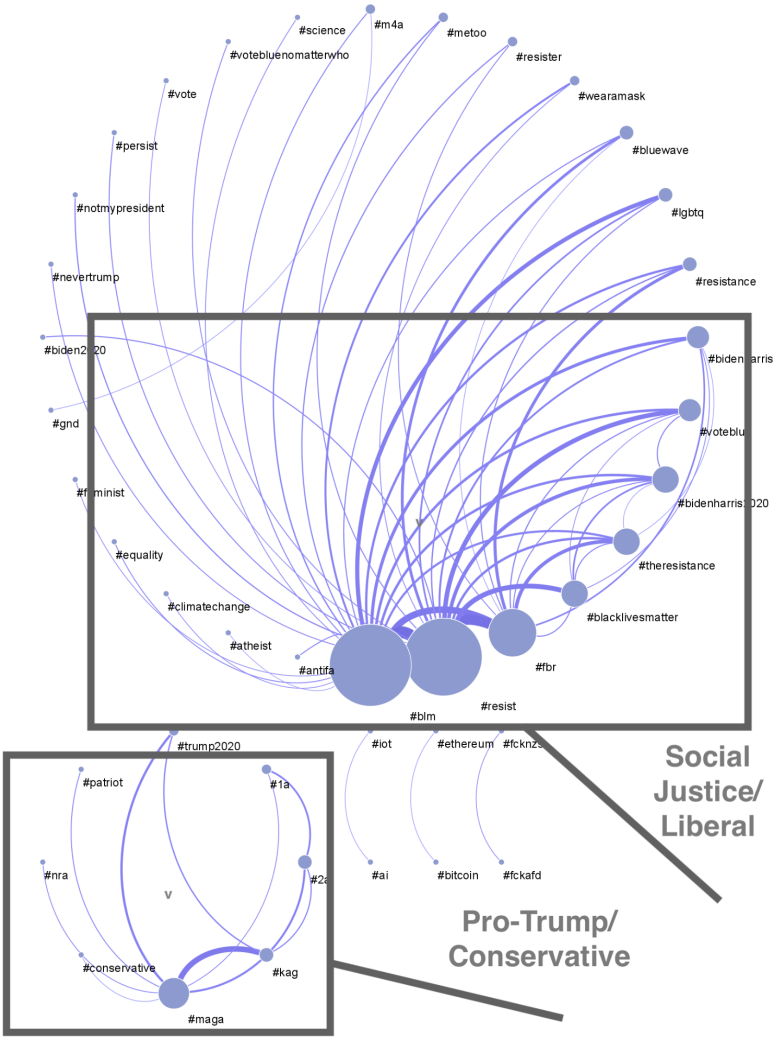

Figure 13: Hashtag Co-occurrences of users sharing fact-checks

Fact-checks have been shown to be effective, even in short-format debunks provided on Twitter (Ecker et al., 2020). Given the polarisation of fact-checking as a field, researchers have suggested that transparency is of great importance in securing users' trust (Humprecht, 2020). In addition, more research is needed into the actual topics with which misinformation and fact-checks can be associated. Bias perceptions have been implicated in the spread of misinformation (Babaei et al., 2021), as research has indicated users rely

considerably on plausibility evaluations in judging the truthfulness of information they encounter (Schwarz & Jalbert, 2019). Recommendations include identifying which items of information may produce the most "bias perceptions" and to address that information first (Babaei et al., 2021).

Co-Occurrences: Further Qualitative Analysis and Future Work

Looking more deeply at the hashtag co-occurrences, we plan to develop a classifier for assigning users to some categories of interest, based on the hashtags they highlight in their user bios on Twitter. Consider Figure 14, in which we have highlighted a portion of Figure XX from our network of hashtags associated with misinforming URLs. We can see several hashtags that are associated with voting for democratic candidates as a form of resistance to the Trump government (#votebluenomatterwho, #votebluetosaveameria), as well as some hashtags that are associated with specific democratic candidates (#votebiden, #bidenharris2020). We also see hashtags that are typically associated with activism undertaken by liberal groups (#blacklivesmatter, #freepalestine, #climatechange), but many of these could be broken down further into specific issues such as anti-racism (#stopasianhate, #blm) or fighting homophobia and transphobia (#lgbtqally, #lgbtq). Understanding which level of granularity will suit the classification task is part of the work we are currently undertaking in the project.

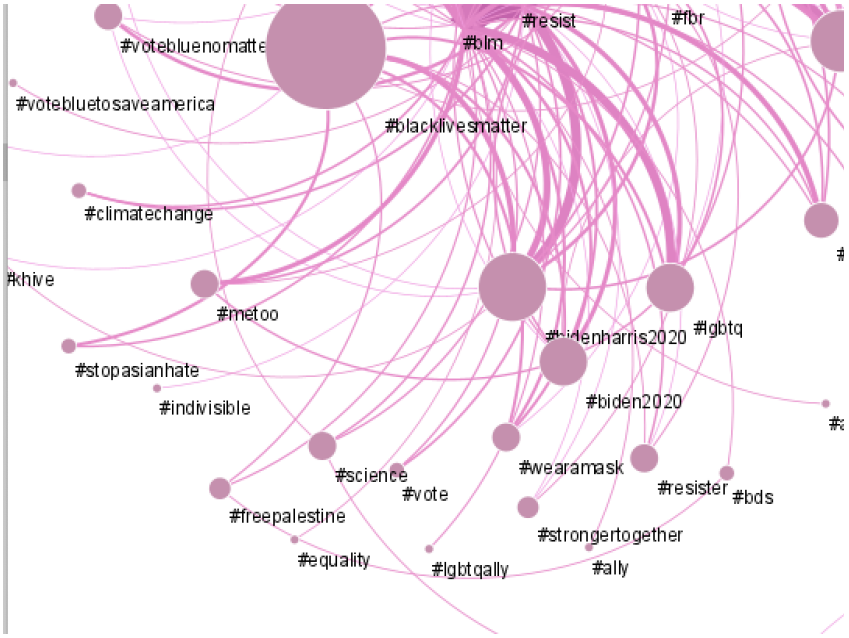

Figure 14: Hashtag Co-occurrences of users sharing misinformation

In Figure 15, we see another portion of the graph looking at users with more conservative hashtags. One can note that many of these hashtags present not only conservative hashtags but conservative hashtags associated specifically with former US president Donald Trump (#buildthewall, #draintheswamp). Another large portion of hashtags relate to patriotic or nationalistic themes (#americafirst, #constitution). At a finer level of granularity, we can even spot specific types of intersectional categories, such as religious conservative (hashtags like #godblessamerica and #prolife, which include both political and religious components). One area of interest is in looking at which types of conservative users on Twitter share misinformation and about which topics. The classical fiscal Republican, for example, is less visible through

these hashtags (with the exception of pro-military hashtags, as support for the military is historically associated with conservative values in the United States). Is this because this type of Republican is less active on Twitter? Less likely to share misinformation? Or do the differences in fact-checking behavior imply different sensemaking processes across the partisan divide? Associating misinformation with particular topics and hashtag co-occurrences might provide some evidence related to these queries.

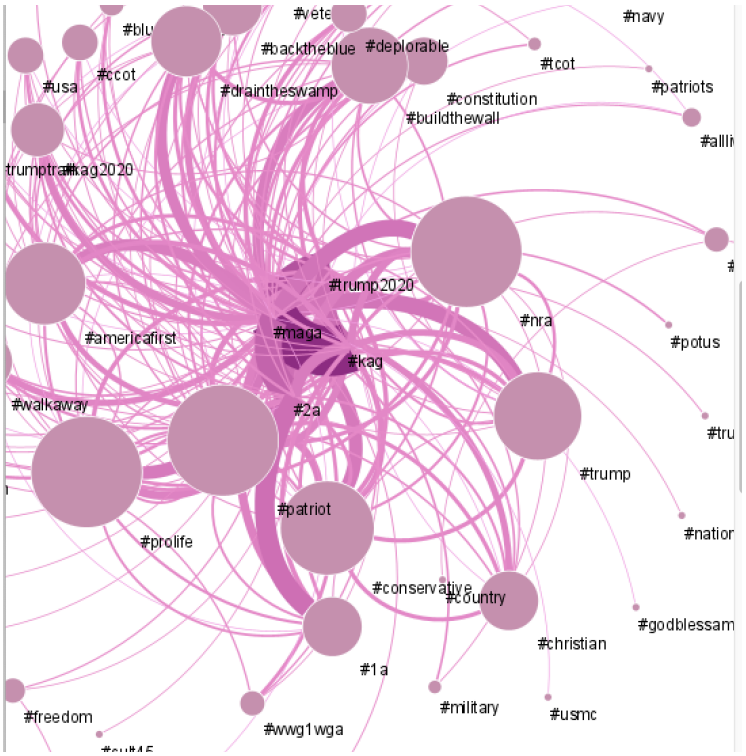

Figure 15: Hashtag Co-occurrences of users sharing misinformation

For example, as a start to the classification process, we did a small analysis of the top hashtags that have more than 100 mentions in our database, which are associated with misinforming posts. We conducted a short annotation exercise to classify these hashtags as belonging to one of 11 macro-categories. For the moment, we disaggregated religion and some social justice categories from partisan categories, to get a sense for the prevalence of purely partisan-related misinformation shares. The chart in Figure 16, shows the prevalence of those categories across misinforming posts. Our analysis concluded that there are many hashtags with ambiguous associations, either because they have been co-opted by opposing groups or because they represent such large themes, it is difficult to identify a clear categorisation. For this reason, we plan to annotate future data according to the prevalence of hashtag co-occurrences. However, we can see from this small analysis that the three groups spotted in the visualisations of the hashtag co-occurrences above in Figure 12 are still visible. More work to refine these categories is currently underway.

Once we have created a set of categories, we plan to enhance our data-driven hashtags through a literature review that includes other hashtags associated with the categories we have defined. The result should be a set of “lexicons” (where hashtags are counted as terms in the lexicon) for each category. The lexicons will be used to classify users by the categories of interest the hashtags in their Twitter bioline

represent. We can then use the classifier to investigate the co-occurence of different categories, related to the topics of misinformation (or fact-checks) that are shared by users in those categories.

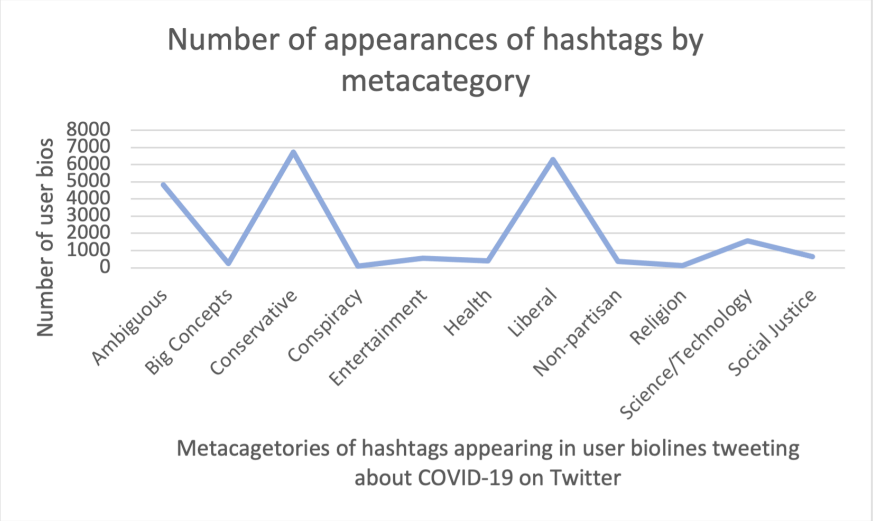

Figure 16: Top hashtag categories of users sharing misinformation

5 Crowd-sourcing Mutual Aid

In the sections above, we focused primarily on the crowd-sourcing of topic information and credibility labels through fact-checkers, and how we might extend and improve this work to understand more about how citizens have dealt with the COVID-19 "infodemic" (Cinelli et al., 2020). In this section, we expand more on our current research into what citizens are doing to respond to some of the challenges triggered by the COVID-19 pandemic and government response. In particular, we focus on mutual aid groups online and our future plans to explore the work of these communities over the course of the COVID-19 pandemic.

Community Disaster Resilience is an ecological, political and social concept that describes a community's capacity or ability to "anticipate, prepare for, respond to, and recover quickly from impacts of disaster" (Mayunga, 2007). In a capital-based model of community disaster resilience (Tierney, 2014), volunteerism and mutual aid can be seen as part of the social capital of a community, "facilitating coordination and cooperation" as well as "access to resources" (Mayunga, 2007). COVID-19 has produced a temporary shared identity, in which a sense of our collective responsibility to one another is linked with our own survival and comfort during the crisis (Drury et al., 2021). As a result, there is a large number of individuals ready to lend a hand to others impacted (sometimes severely) by the pandemic, not as an act of charity, but as an act of solidarity[17].

A community-managed Covid Mutual Aid wiki[18] has tracked nearly 6000 mutual aid groups globally. They are found in urban and rural areas, and more are added to this community resource each day, according to the wiki's website. A regional website for COVID mutual aid groups just in the UK[19] reports more than 4000

[17] https://www.democracynow.org/2020/3/20/coronavirus_community_response_mutual_aid
[18] https://mutualaid.wiki/
[19] https://covidmutualaid.org/local-groups/

groups in the UK alone. Evidence suggests that mutual aid groups have been able to organise themselves and respond quickly, sometimes weeks ahead of any government provision of services or risk mitigation strategies.[20] A rapid survey conducted by the Community Support and Mutual Aid research team at the University of Sussex (Mao et al., 2021) reported that mutual aid groups have shifted their focus during the course of the pandemic, addressing problems associated with shielding and lockdowns (running errands, walking dogs, filling prescriptions) at the beginning of the pandemic and providing more emotional and social support as partial lockdowns continued. In the later stages of the pandemic (the report covers up to October 2020), mutual aid groups began to focus on some of the groups most marginalised by the pandemic, such as the homeless or those threatened by domestic abuse. Evidence suggests that some of these groups grow out of existing volunteer efforts in a community and re-emerge during successive crises, suggesting that the sustainability of such groups is an important feature of Community Disaster Resilience.

In April 2020, we conducted our own initial qualitative study into a local community aid group in the UK to understand more about who was taking part in such groups, how they communicated with each other and what kinds of help were being offered or requested. The community we studied has approximately 15,000 residents and 6,200 homes, meaning that households are likely to be smaller. The area is also above the national average in work-based earnings. For this reason, we classify the area as middle class to affluent. On one hand, the choice of an affluent area allows us to surface more consequences that are directly related to COVID-19, than to other potential factors. On the other, it does not allow us to see how the nation's most marginalised people have been dealing with the additional burdens that COVID-19 has produced. We discuss these limitations later in this deliverable and our plans for future work in this area.

We posted our survey in the local facebook group and asked residents involved with the group to share their experiences. We had 30 respondents. Of our respondents, nearly all were white (n=28), female (n=28) and over the age of 40 (n=24). During the lockdown phases of the pandemic, women were more likely to be furloughed and to take on additional childcare in the home after school closures,[21] which may have led to the situation that more women were available to take part in mutual aid schemes. Of course, this could also be a self-selection bias in our survey.

Most were sharing their home with 3 or fewer other individuals (n=21), which reflects the resident-to-home ratio reported for the area. This means that, during the initial phases of the pandemic, many individuals had little opportunity for physical social contact. This may have also encouraged participation in local mutual aid schemes. Time may have been an additional factor, due to lockdowns and the furlough scheme. In our study, most participants were retired, or receiving their same income as before, due to continuing working, working from home or being furloughed at their same salary. The remaining respondents were unable to work due to unemployment (n=2), disability (n=1) or lockdown (n=1), or working at a reduced salary/in a different role (n=2). 50% of the respondents had enough money to support themselves through the COVID crisis, with additional savings. 33.7% reported that they had enough money to cover their basics, but no savings. Only two of the respondents reported that they were struggling to make ends meet, while two respondents reported no change.

[20] Rapid Research into COVID-19, https://www.cso.scot.nhs.uk/wp-content/uploads/COVCGU2006.pdf
[21] Office of National Statistics UK, https://tinyurl.com/2fcsvxkw

More than half of the participants said that they were coping through the crisis and had a handle on their mental health (n=17). However, 33% (n=11) reported problems with mental health and 2 additional respondents reported regular mental health concerns. 43.3% of participants reported that they have needed help during COVID. 56.7% reported that they did not need any assistance. 80% had offered help to others directly and 48.3% of participants reported others in their household offering assistance as well. By far, the greatest percentage of assistance needed was around grocery shopping and picking up prescriptions. This is followed by needing temporary financial assistance or help setting up/managing online orders and relationships. Those offering help reported similarly, with running errands for those who could not leave home at the top of the list. However, more than half of participants also reported offering emotional support. People were offering assistance every week, sometimes multiple days per week, to friends, family, but also complete strangers. This is despite feeling worried about contracting the virus and feeling at risk. On a scale of 1 to 5, where 1 is no risk and 5 is high risk, more than half of participants rated their risk of contracting the virus at 3 or higher. On that same scale, more than half of participants rated their worry about coming into contact with the virus at 4 or higher. While our data is limited and cannot be extrapolated to understand all mutual aid communities, our findings are in line with those of the Office for National Statistics during that time, with regard to financial consequences, concerns about the virus and perceived mental health disturbances.[22] They also mirror findings by other researchers focusing on mutual aid at this time (as described above): mutual aid appears to be successful in areas with high social capital. Social capital is linked to having (existing) networks of individuals willing to coordinate a local response, provide assistance where needed, and respond to the changing landscape of the crisis.

One year on, we contacted the administrators of the group to ask them some general questions about local community groups with which they cooperated and to estimate the level of need within the community. We learned that the local group was in touch with local hospitals, citizens' advice networks, food banks and the town council, indicating good integration with other more formalised services. In addition, administrators estimated the size of their mutual aid community in the hundreds, with a ratio of those providing help to those needing help at 1:3. They reported a success rate (being able to fulfil needs requests) at 100%, indicating a high degree of efficacy even one year on. When asked about other features of the pandemic that influenced their work or the conditions of their work, administrators reported supply chain disruptions (such as panic buying or transport link disruption) and misinformation about the pandemic as continual challenges.

Planned Research

To analyse the activities of such groups at scale, and to relate their work with other data we have about food shortages, government guidelines, characteristics of the pandemic, etc., we can use some of the methods described above to:

1. Track requests for assistance and offerings of assistance over time to see how and when mutual aid groups were most effective during COVID-19.
2. Investigate the role and behaviours of gatekeepers and boundary spanners within Mutual Aid groups.

[22] Coronavirus and social impacts, https://tinyurl.com/274sbe46

3. Understand the connection of current COVID-19 mutual aid groups to groups that emerged during previous crises, to understand more about the sustainability and activation of such groups.

For this we need to be able to create a taxonomy of needs and offerings such that they can be identified and classified automatically from social media data. At the moment we are investigating available data and local partnerships to understand what is already known about mutual aid groups and their activities during COVID. We also need to identify the names and locations of such groups, and their social media accounts. By looking at the date an account was created, we can start to get a sense for whether or not such groups were active before COVID-19, or were created specifically for this purpose. We can also observe the follower networks of such groups, and identify topics of interest through hashtags and key word searches, to get a sense for who is contributing to mutual aid groups. If Bourdieu's proposition is correct, we should see a wide diversity of topics of interests and hashtags in user bios, indicating the mutual interest in cooperation through crisis.

6 Conclusion

Task 4.2 deals with the task of automatically and accurately processing crowd information to enhance the situational awareness of citizens during a crisis. In this deliverable, we have presented our approach to analyzing public COVID-19 data sharing habits on social media, and introduce the framework of mutual aid to ground our future research in this area. We describe several approaches that we believe will help us to understand more about how the public respond to informational needs about COVID-19, and the material needs arising from different government interventions of the situation on the ground.

To understand informational needs, our Fact-Checking Observatory (FCO) generates weekly reports on the spread of misinformation topics and fact-checks. As the goal is to do this with no human input, we have generated an automatic classification system for misinformation topics, although the **confusion rate remains high**. The reasons for this include the model being limited to one label per post, when misinformation narratives are frequently blended together. In the future we may switch to an approach allowing multiple topic classes per tweet.

From the FCO reports, we obtained a list of hashtags present in user biographies, which we then analyzed to gain further insight into who specifically was spreading misinformation or fact-checks. Though the metacategories we identified were diverse, network mapping revealed two polarized camps similar to those in the US political arena, with a third group centering around technological interests such as AI. Few bridges between networks were visible. **While Right- and Left-wing coded accounts both shared misinformation, fact-checks were far less prevalent in Right-wing coded accounts.** Classifying these users into more granular categories remains a major goal. We can use this analysis to understand the influence of partisanship or ideology on the provision of factual information (and potentially goods and services) during the COVID-19 crisis.

The next step in our research, as we have described, is to follow the activities of mutual aid groups and their followers on Twitter, tracking their exposure to misinformation and verified information wherever possible. Mutual aid groups have used social media to respond to the needs of the public in a timely manner, which suggests they are following the media and information about the crisis. To understand more about the activities and motivations of mutual aid groups, we report on a single case study we conducted in mid-2020. In line with the existing literature, this study indicated that mutual aid groups may build on existing community groups and appear more effective in places with high social capital. Follow-up one year afterwards showed that **the group was 100% effective in responding to material requests for aid, but supply chain disruptions, panic buying, and misinformation remain recurring challenges.** Understanding more about the sustainability of mutual aid groups, when and where they arise, as well as the influence of different types of information in their networks are important topics for future research.

7 References

Aguilar-Gallegos, N., Romero-García, L. E., Martínez-González, E. G., García-Sánchez, E. I., & Aguilar-Ávila, J. (2020). Dataset on dynamics of Coronavirus on Twitter. *Data in Brief*, *30*, 105684. https://doi.org/10.1016/j.dib.2020.105684

Amazeen, M. A., Vargo, C. J., & Hopp, T. (2019). Reinforcing attitudes in a gatewatching news era: Individual-level antecedents to sharing fact-checks on social media. *Communication Monographs*, *86*(1), 112–132. https://doi.org/10.1080/03637751.2018.1521984

Babaei, M., Kulshrestha, J., Chakraborty, A., Redmiles, E. M., Cha, M., & Gummadi, K. P. (2021). Analyzing Biases in Perception of Truth in News Stories and Their Implications for Fact Checking. *IEEE Transactions on Computational Social Systems*, 1–12. https://doi.org/10.1109/TCSS.2021.3096038

Balasubramanian, V. N., Ho, S.-S., & Vovk, V. (Eds.). (2014). Conformal Prediction for Reliable Machine Learning. In *Conformal Prediction for Reliable Machine Learning* (p. i). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-398537-8.00014-6

Brandt, J., Buckingham, K., Buntain, C., Anderson, W., Ray, S., Pool, J.-R., & Ferrari, N. (2020). Identifying social media user demographics and topic diversity with computational social science: A case study of a major international policy forum. *Journal of Computational Social Science*, *3*(1), 167–188. https://doi.org/10.1007/s42001-019-00061-9

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media infodemic. *Scientific Reports*, *10*(1), 16598. https://doi.org/10.1038/s41598-020-73510-5

Clarke, L. (2021). Covid-19: Who fact checks health and science on Facebook? *BMJ*, *373*, n1170. https://doi.org/10.1136/bmj.n1170

Demographic Inference and Representative Population Estimates from Multilingual Social Media Data | The World Wide Web Conference. (n.d.). Retrieved September 16, 2021, from https://dl.acm.org/doi/10.1145/3308558.3313684

Drury, J., Carter, H., Ntontis, E., & Guven, S. T. (2021). Public behaviour in response to the COVID-19 pandemic: Understanding the role of group processes. *BJPsych Open*, *7*(1), e11. https://doi.org/10.1192/bjo.2020.139

Ecker, U. K. H., O'Reilly, Z., Reid, J. S., & Chang, E. P. (2020). The effectiveness of short-format refutational fact-checks. *British Journal of Psychology*, *111*(1), 36–54. https://doi.org/10.1111/bjop.12383

Gilardi, F., Gessler, T., Kubli, M., & Müller, S. (2021). Social Media and Policy Responses to the COVID-19 Pandemic in Switzerland. *Swiss Political Science Review*, *27*(2), 243–256. https://doi.org/10.1111/spsr.12458

Guess, A., & Nyhan, B. (n.d.). *Selective Exposure to Misinformation: Evidence from the consumption of fake news during the 2016 U.S. presidential campaign*. 49.

Humprecht, E. (2020). How Do They Debunk "Fake News"? A Cross-National Comparison of Transparency in Fact Checks. *Digital Journalism*, *8*(3), 310–327. https://doi.org/10.1080/21670811.2019.1691031

Karami, A., Kadari, R. R., Panati, L., Nooli, S. P., Bheemreddy, H., & Bozorgi, P. (2021). Analysis of Geotagging Behavior: Do Geotagged Users Represent the Twitter Population? *ISPRS International Journal of Geo-Information*, *10*(6), 373. https://doi.org/10.3390/ijgi10060373

Karisani, N., & Karisani, P. (2020). Mining Coronavirus (COVID-19) Posts in Social Media. *ArXiv:2004.06778 [Cs, Stat]*. http://arxiv.org/abs/2004.06778

Khare, P., Burel, G., Maynard, D., & Alani, H. (2018). Cross-Lingual Classification of Crisis Data. In D. Vrandečić, K. Bontcheva, M. C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L.-A. Kaffee, & E. Simperl (Eds.), *Lecture Notes in Computer Science* (Vol. 11136, pp. 617–633). Springer. http://oro.open.ac.uk/57253/

Lyu, H., Chen, L., Wang, Y., & Luo, J. (2020). Sense and Sensibility: Characterizing Social Media Users Regarding the Use of Controversial Terms for COVID-19. *IEEE Transactions on Big Data*, 1–1. https://doi.org/10.1109/TBDATA.2020.2996401

Mao, G., Fernandes-Jesus, M., Ntontis, E., & Drury, J. (2021). What have we learned about COVID-19

volunteering in the UK? A rapid review of the literature. *BMC Public Health*, *21*(1), 1470. https://doi.org/10.1186/s12889-021-11390-8

Mayunga, J. S. (2007). *Understanding and Applying the Concept of Community Disaster Resilience: A capital-based approach*. 16.

Naseem, U., Razzak, I., Khushi, M., Eklund, P. W., & Kim, J. (2021). COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis. *IEEE Transactions on Computational Social Systems*, *8*(4), 1003–1015. https://doi.org/10.1109/TCSS.2021.3051189

Nyhan, B., & Reifler, J. (n.d.). *Estimating Fact-checking's E ects*. 18.

Oktay, H., Firat, A., & Ertem, Z. (2014). *Demographic Breakdown of Twitter Users: An analysis based on names*.

Raamkumar, A. S., Tan, S. G., & Wee, H. L. (2020). Use of Health Belief Model–Based Deep Learning Classifiers for COVID-19 Social Media Content to Examine Public Perceptions of Physical Distancing: Model Development and Case Study. *JMIR Public Health and Surveillance*, *6*(3), e20493. https://doi.org/10.2196/20493

Rich, T. S., Milden, I., & Wagner, M. T. (2020). Research note: Does the public support fact-checking social media? It depends who and how you ask. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-46

Schwarz, N., & Jalbert, M. (2019). *When (Fake) News Feels True: Intuitions of Truth and the Acceptance and Correction of Misinformation*.

Tekumalla, R., & Banda, J. M. (2020). *Characterization of Potential Drug Treatments for COVID-19 using Social Media Data and Machine Learning*. 7.

Tierney, K. (n.d.). *Chapter 8. Adaptive Resilience in the Face of Disasters*. Retrieved September 16, 2021, from https://www.degruyter.com/document/doi/10.1515/9780804791403-009/html

Vashisth, P., & Meehan, K. (2020). Gender Classification using Twitter Text Data. *2020 31st Irish Signals and Systems Conference (ISSC)*, 1–6. https://doi.org/10.1109/ISSC49989.2020.9180161